UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ISABELLA GRZECZECZEN GONCALVES

UM ESTUDO SOBRE A APLICABILIDADE DE BIG DATA NA REDE SOCIAL TWITTER

TRABALHO DE DIPLOMAÇÃO

MEDIANEIRA
2015

ISABELLA GRZECZECZEN GONÇALVES

UM ESTUDO SOBRE A APLICABILIDADE DE BIG DATA NA REDE SOCIAL TWITTER

Trabalho de Diplomação apresentado à disciplina de Trabalho de Diplomação, do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas – COADS – da Universidade Tecnológica Federal do Paraná – UTFPR, como requisito parcial para obtenção do título de Tecnólogo.

Orientador: Prof. *MSc* Juliano Rodrigo Lamb.

MEDIANEIRA

2015

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Departamento Acadêmico de Computação
Curso de Tecnologia em Análise e Desenvolvimento de Sistemas

TERMO DE APROVAÇÃO

UM ESTUDO SOBRE A APLICABILIDADE DE *BIG DATA* NA REDE SOCIAL TWITTER

Por

ISABELLA GRZECZECZEN GONÇALVES

Este Trabalho de Diplomação (TD) foi apresentado à 13h30min do dia 10 de Junho de 2015, como requisito parcial para a obtenção do título de Tecnólogo no Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O acadêmico foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. MSc Juliano Rodrigo Lamb UTFPR – Câmpus Medianeira (Orientador)	Prof. Márcio Angelo Matté UTFPR – Câmpus Medianeira (Convidado)
Prof. Dr. Everton Coimbra de Araújo UTFPR – Câmpus Medianeira (Convidado)	Prof. MSc Juliano Rodrigo Lamb UTFPR – Câmpus Medianeira (Responsável pelas atividades de TCC)

A folha de aprovação assinada encontra-se na Coordenação do Curso.

AGRADECIMENTOS

É certo de que estas linhas não serão suficientes para mencionar todas as pessoas que fizeram parte desta importante fase da minha vida. Porém elas podem estar certas de que fazem parte do meu pensamento e de minha gratidão.

Primeiramente agradeço a Deus por ter me dado forças e iluminando meu caminho para que pudesse concluir mais uma etapa da minha vida;

A minha querida mãe, que nunca mediu esforços para realização de meus sonhos;

Agradeço também aos professores MSc Juliano Rodrigo Lamb e MSc Alan Gavioli, por toda dedicação, orientação e apoio durante todo o trabalho desenvolvido;

A todos os professores do curso de Análise e Desenvolvimento de Sistemas, pela paciência, dedicação e ensinamentos disponibilizados nas aulas, cada um de forma especial contribuiu para a conclusão desse trabalho e consequentemente para minha formação acadêmica;

Aos amigos que fiz durante o curso, pela amizade que construímos, por todos os momentos que passamos durante esta trajetória;

Por fim, gostaria de agradecer aos meus amigos e familiares, pelo carinho e apoio, a todos que contribuíram direta ou indiretamente para que esse trabalho fosse realizado. Muito obrigada!

RESUMO

GONCALVES, Isabella Grzeczeczen. UM ESTUDO SOBRE A APLICABILIDADE DE *BIG DATA* NA REDE SOCIAL TWITTER. 2015. 62 f. Trabalho de Conclusão do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas. Universidade Tecnológica Federal do Paraná. Medianeira, 2015.

O presente trabalho é embasado teoricamente em estudos sobre a aplicação de *Big Data* nas redes social *Twitter*. O trabalho teve como objetivo a criação de um estudo sobre as ferramentas utilizadas para a mineração dos dados por meio da análise de sentimento. Teve como ferramentas principais o *Apache Hadoop* e o *Hortonworks Data Plataform* que foram imprescindíveis para a análise de *Big Data*. Este estudo possibilitou um programador por meio de *Hadoop* pudesse extrair dados de sentimento do *Twitter* para analisar o desempenho de um lançamento de um produto. Para isso, foi necessário fazer o download e extrair os arquivos de sentimento criados por meio da coleta do Apache Flume, agregando grandes quantidades de dados de fluxo para o *Hadoop Distributed File System (*HDFS). Em seguida, foi feito o carregamento dos arquivos de dados do Twitter criados para a *Hortonworks Sandbox*, o *singlenode cluster Hadoop* rodado na Máquina Virtual. Foi utilizado o *HCatalog* para construir uma visão relacional dos dados onde, em seguida foi feita a cópia e execução do script *Hive* para a *Sandbox* para o possível refinamento dos dados brutos e consulta destes dados. Por fim foi possível importar e acessar esses dados refinados com a utilização do Microsoft Excel, e a visualização dos dados de sentimento usando o *Excel Power View*.

Palavras-chaves: *Apache Hadoop*, análise de sentimentos, mineração de dados.

ABSTRACT

GONCALVES, Isabella Grzeczeczen. UM ESTUDO SOBRE A APLICABILIDADE DE *BIG DATA* NA REDE SOCIAL TWITTER. 2015. 62 f. Trabalho de Conclusão do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas. Universidade Tecnológica Federal do Paraná. Medianeira, 2015.

This study is theoretically grounded in research on the application of *Big Data* on social network *Twitter*. The work aims to create a study of the tools used for *Data Mining* through sentiment analysis. And had as main tools the Apache Hadoop and the Hortonworks Data Platform that were essential for the analysis of *Big Data*. This study enabled a programmer through Hadoop could extract the Twitter sentiment data to analyze the performance of a release of a particular movie. For this, you need to download and extract the feeling of files created by collecting the Apache Flume, adding large amounts of flow data to the Hadoop Distributed File System (HDFS). Then loading the Twitter data files created for Hortonworks Sandbox was done, the Hadoop cluster singlenode shot in Virtual Machine. It was used the HCatalog to build a relational view of the data where then the copying is done and execution of the Hive script for the Sandbox for possible refinement of raw data and query this data. Finally we were able to download and access these refined data using Microsoft Excel, and the sense of data displayed using Excel Power View.

Keywords: Apache Hadoop, sentiment analysis, *Data Mining*.

LISTA DE SIGLAS

AI	*Artificial Intelligence*
API	*Application Programming Interface*
BI	*Business inteligence*
DM	*Data Mining*
DW	*Data Warehouse*
HDFS	*Hadoop Distributed File System*
IDE	*Integrated Development Environment*
ODBC	*Open Database Connectivity*
OLAP	*Online Analytical Processing*
PDF	*Portable Document Format*
SQL	*Structured Query Language*
TI	*Tecnologia Da Informação*
VM	*Virtual Machine*

LISTA DE FIGURAS

Figura 1 - Big Data e os 5Vs. 16
Figura 2 - Características *Big Data* x *Data Mining* 17
Figura 3 - Funcionamento do BI 19
Figura 4 - Arquitetura Hadoop. 21
Figura 5 - Estrutura Pentaho. 23
Figura 6 - Computação em Nuvem 26
Figura 7 - Arquitetura MapReduce 28
Figura 8 - Funcionamento do HDFS 30
Figura 9 - Hadoop e suas Características 31
Figura 10 - Hortonsworks Data Platform 32
Figura 11 - Iniciando a *Hortonworks Sandbox* na *Virtual Box*. 33
Figura 12 - *Console Sandbox* 34
Figura 13 - Página Inicial Sandbox 34
Figura 14- Amostra de dados de sentimento via Twitter 35
Figura 15 - Configuração Twitter 37
Figura 16- Instalando o Flume na Sandbox 37
Figura 17 - Flume em execução 38
Figura 18 - Carregamento do arquivo .zip upload 38
Figura 19 - Arquivo extraído no diretório 39
Figura 20 - Transferência de arquivos com WinSCP 39
Figura 21 - Login no Console Sandbox 40
Figura 22 - Execução do script Hive 40
Figura 23 - Tabela tweets_raw 41
Figura 24 - Tabela Dictionary 42

Figura 25 - Tabela Tweetsbi 42
Figura 26 - Listagem das tabelas adicionadas ao diretório Sandbox. 43
Figura 27 - Tabelas na aba HCatalog 43
Figura 28 - Query Results 44
Figura 29 - Microsoft Query 45
Figura 30 - Escolha da Fonte de Dados 46
Figura 31 - Escolher colunas 46
Figura 32 - Dados Importados 47
Figura 33 - Inserir Power View 48
Figura 34- Visualizar dados com o Mapa 48
Figura 35 - Visualização do Mapa Global: Sentimento x País 49
Figura 36 - Contagem de tweets por país e sentimento 50
Figura 37 - Contagem de sentimentos neutros no Brasil 51
Figura 38 - Contagem de Sentimentos Positivos nos Estados Unidos 51
Figura 39 - Contagem de sentimentos negativos no Canadá. 52

SUMÁRIO

1 INTRODUÇÃO 9
1.1 OBJETIVO GERAL 10
1.2 OBJETIVOS ESPECÍFICOS 10
1.3 JUSTIFICATIVA 10
2 FUNDAMENTAÇÃO TEÓRICA 12
2.1 REDES SOCIAIS 12
2.1.1 Mídias Sociais 12
2.2 BIG DATA 14
2.3 BIG DATA E O CONCEITO DOS 5 VS 15
2.4 BIG DATA E DATA MINING 17
2.5 DATA MINING E BUSISNESS INTELIGENCE 18
2.6 BIG DATA E BUSINESS INTELLIGENCE 18
2.7 TECNOLOGIAS E FERRAMENTAS ASSOCIADAS À BIG DATA 20
2.7.1 Apache Hadoop 20
2.7.2 Jasper soft BI Suite 22
2.7.3 Pentaho Business Analytics 22
2.7.4 Karmasphere Studio 24
2.7.5 Talend Open Studio 25
2.7.6 Skytree Servidor 25
2.7.7 Tableau Software 25
2.7.8 Splunk 26
2.8 CLOUD COMPUTING E BIG DATA 26
3 MATERIAL E MÉTODOS 27
3.1 MAPREDUCE 27
3.2 HADOOP DISTRIBUTED FILE SYSTEM (HDFS) 30
3.3 HORTONWORKS DATA PLATAFORM 32
3.3.1 Hortonworks Sandbox 33
3.4 ESCOPO DO ESTUDO 35
3.4.1 Obter os dados de sentimento por meio do Apache Flume 36
3.4.2 Refinando os dados com MapReduce 39

3.4.3 Classificação de sentimento em positivo, negativo ou neutro. 41
4 RESULTADOS E DISCUSSÃO 45
4.1 IMPORTAÇÃO DOS DADOS NO EXCEL 45
4.2 VISUALIZAÇÃO DOS RESULTADOS 47
4.3 APLICAÇÃO DOS RESULTADOS 52
5 CONSIDERAÇÕES FINAIS 54
5.1 CONCLUSÃO 54
5.2 TRABALHOS FUTUROS/CONTINUAÇÃO DO TRABALHO 55
6 REFERÊNCIAS BIBLIOGRÁFICAS 56

1 INTRODUÇÃO

Uma nova era tecnológica está em desenvolvimento: um cenário em que o volume de dados aumenta a cada instante e em que soluções podem ser encontradas em informações que, aparentemente, não possuiriam relação alguma. Além disso, dados na casa dos *zetabytes* já podem ser considerados algo muito próximo de ser alcançado (WHITE, 2010).

A sociedade tornou-se conectada tendo participado do envio e recebimento de dados por meio da Internet. As redes sociais, mensagens de texto, envio de vídeos, registro de transações de compras, cliques em sites e sensores em dispositivos, entre outros, fazem parte de um grande banco de dados não estruturado que está disponível para empresas que querem investir em tomada de decisões. Grande parte desses dados é criada pela própria sociedade numa escala que vem chamando atenção (CISCO, 2013). Analisar como estes dados permeiam os níveis estratégicos das organizações é fundamental para o sucesso no mundo dos negócios. As diversas características das informações, assim como as formas de classificá-las, são determinantes na escolha das ferramentas que serão utilizadas no seu tratamento (TAURION, 2012).

A explosão de dados nas redes sociais tem se tornado uma verdadeira mina de ouro para as empresas das áreas de marketing e vendas, desempenho operacional e financeiro, e inovação. Consumidores se tornaram importantes formadores de opinião, compartilhando publicamente, seus pensamentos com relação a produtos e serviços. Para analisar estes dados importantes, estas empresas têm utilizado a análise de sentimentos ou análise de opinião para entender a preferência de seus clientes, tendências e reconhecimento da sua marca perante o mercado. (BROWN, 2012).

O tamanho da empresa, o volume, a velocidade, a veracidade e a variedade dos dados são insumos para um novo mundo de oportunidades. (HENRIQUES, 2013). Este trabalho busca discorrer sobre os conceitos e ferramentas de *Big Data*, as estratégias de aplicação, benefícios e desafios, e também, possibilidades de ganhos reais que se podem obter com a utilização deste importante recurso que é a mineração de dados.

1.1 OBJETIVO GERAL

Aplicar os conceitos e ferramentas de *Big Data* na mineração de dados na rede social *Twitter*, por meio da prática de análise de sentimento, que visa identificar o sentimento ou opinião das pessoas expressa de forma escrita ou falada baseado no conteúdo disponível na *Web.*

1.2 OBJETIVOS ESPECÍFICOS

- Elaborar uma contextualização sobre *Big Data* e tecnologias associadas;
- Realizar um levantamento sobre a aplicação de *Big Data* nas redes sociais;
- Realizar um estudo sobre a análise de sentimento na rede social *Twitter*;
- Identificar vantagens e desvantagens da utilização de *Big Data* nas redes sociais.

1.3 JUSTIFICATIVA

Tendo em vista toda essa evolução no mundo dos dados, surge a pergunta: o que pode ser feito com toda essa quantidade de informação disponível na *Web*? No mundo dos negócios, decisões que antes eram baseadas em suposições, ou em modelos construídos por especialistas, podem ser feitas com base nos dados coletados. (BROWN, 2011).

McAfee e Brynjolfsson (2012) conduziram estudos que levaram à conclusão de que as empresas que efetivamente utilizam *Big Data* são 5% mais produtivas e 6% mais lucrativas que seus competidores – na atualidade esses números são um poderoso argumento em prol da utilização dessa abordagem.

Segundo Brown (2011), *Big Data* permite que as empresas criem *frameworks* para testar hipóteses que possam auxiliar na tomada de decisão, gerando assim, a possibilidade de uma tomada de decisão diferente, possibilitando diferenciar um simples grupo de eventos daqueles que realmente possam ter uma ligação.

Brynjolfsson (2012) afirma que os empresários necessitam se adaptar a este novo modelo de gestão o mais rápido possível, pois as soluções de *Big Data* têm um potencial muito maior que as soluções de análise tradicional para beneficiar as empresas e oferecer vantagens para quem já as utiliza.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta o referencial teórico das tecnologias e ferramentas utilizadas em *Big Data* e também suas aplicações.

2.1 REDES SOCIAIS

Redes sociais são estruturas virtuais formadas por pessoas e/ou organizações, conectadas por um ou vários tipos de relações, que possuem valores e objetivos em comum na Internet. (LEITE, 2014).

> Estrutura social refere-se à colocação e à posição de indivíduos e de grupos dentro desse sistema. O agrupamento de indivíduos, de acordo com as posições que resultam dos padrões essenciais de relações de obrigação, constitui a estrutura social de uma sociedade. E são virtuais, obviamente, porque se refere às comunicações via Internet que extrapolam as quatro paredes convencionais. (BROWN, 2011).

Segundo Cardozo (2009), redes sociais representam um conjunto de participantes, que unem ideias em torno de valores e interesses em comum. O ponto central é a valorização dos canais informais e das relações, em consideração das estruturas hierárquicas.

As redes sociais fazem parte das mídias sociais, que é o conteúdo de forma não centralizada, em que não há o controle editorial de grandes grupos. Pode-se chamar de produção de muitos para muitos. (LEITE, 2014).

2.1.1 Mídias Sociais

Mídia social trata-se de tecnologias e práticas *online* que pessoas usam para compartilhar opiniões, perspectivas e ideias, podendo se expressar de diferentes formas como: textos, áudio, imagens e vídeo. Com o crescimento dessas mídias tornou-se mais fácil encontrar amigos, compartilhar opiniões e obter informações, ou seja, todos participam na construção de uma comunidade virtual. (SOLIS, 2007).

Segundo Manovich (2011) existem diversos tipos de mídias sociais com as mais diferentes categorias. As mais conhecidas são:

- Mídia de colaboração: relacionada às redes sociais com a interação de diferentes usuários que compartilham informações em comum. Como exemplo destacam-se a Wikipédia[1], Yelp[2] e Digg[3].
- Mídia de comunicação: está relacionada à conversação entre pessoas, por meio de comentários e compartilhamento conteúdo. Como exemplo, é possível mencionar *WordPress*[4]*, Twitter*[5]*, Facebook*[6] *e GoogleGroups*[7].
- Multimídia: refere-se aos componentes audiovisuais que ficam além do texto puro, como fotos, vídeos e músicas. Alguns exemplos dessas mídias, Flickr[6], YouTube[7], JustinTV[8] e Lastfm9.
- Entretenimento: refere-se um mundo virtual favorecendo o desenvolvimento de ambientes focados em games online ou ainda atividades que podem ser transformadas em algum tipo de competição. Como exemplo podem ser citados o Second Life[8] e TvTag[9].

Estas plataformas se tornam populares porque são evidenciadas por meio da capacidade que possuem de produzir enormes volumes de conteúdo. A monitoração destas mídias sociais tornou-se uma responsabilidade de *Big Data*, onde se precisa tratar um volume enorme de dados, levando em consideração que essas mídias possuem características diferentes em relação à estrutura, dinâmica, uso e modelagem e necessidade de velocidade em seu tratamento para que diferentes análises sejam viáveis. (SOLIS, 2007).

[1] Disponível em: www.wikipedia.com.br
[2] Disponível em: www.yelp.com.br
[3] Disponível em: www.digg.com
[4] Disponível em: https://br.wordpress.org
[5] Disponível em: www.twitter.com
[6] Disponível em: www.facebook.com
[7] Disponível em: https://groups.google.com
[8] Disponível em : www.secondlife.com
[9] Disponível em: www.tvtag.com

2.2 BIG DATA

Costa et al. (2012) conceituam *Big Data* como um enorme volume de dados que pela inexistência, no cenário atual, de ferramentas específicas de manipulação, exige um maior esforço para o seu gerenciamento, visto que tal gerenciamento é realizado por aplicações moldadas para trabalhar com um volume menor de informações. Desta forma, para classificar um determinado bloco de dados como *Big Data*, é necessário observar se a ferramenta destinada para sua manipulação conseguirá processar a demanda de informações, de modo a alcançar satisfatoriamente o objetivo proposto.

> A quantidade de dados armazenados sobre os mais diversos assuntos cresce a uma taxa muito elevada. Empresas capturam volumes de bytes de informações sobre seus clientes, fornecedores e funcionários, e milhões de sensores conectados estão sendo inseridos no mundo físico em aparelhos como celulares e automóveis, percebendo, criando e comunicando dados. Indivíduos com *smartphones* e em *sites* de redes sociais continuarão incrementando crescimento exponencial. *Big Data* – grandes repositórios de dados que podem ser capturados, comunicados, agregados, armazenados e analisados – é parte de cada setor e função da economia global (MANYIKA et al, 2011).

Há alguns anos, guardar e acessar com eficiência um volume relativamente pequeno de informações exigia um investimento pesado em *hardware*, com o investimento financeiro considerável, além de a instalação demorar muito tempo. Nos anos 1990, estocar um *gigabyte* custava aproximadamente 1000 dólares para as empresas. No momento presente, custam seis centavos apenas. (FEIJÓ, 2013).

As definições existentes na literatura para o *Big Data* convergem para os seguintes fatos: a utilização de diferentes fontes, tipos de dados e características que se refere ao volume, variedade e velocidade (Manyka et al. (2011); IBM; Begolli e Hovey (2012); McAfee e Brynjolfisson (2012). Os autores Zikopoulos e Eaton (2011), acrescentam o atributo veracidade. Há menos de 10 anos, usava-se muito o termo *Giga-Bytes* e rapidamente mudou-se para *Terabaytes* e atualmente se fala em *Pentabytes* e *Exabytes*. Desta maneira é possível observar como o volume de dados no mundo cresce rapidamente.

A análise das informações quando bem executada, pode levar ao conhecimento de padrões de comportamento, significados antes ocultos e ajudar a prever tendências de consumo. (FEIJÓ, 2013).

Uma possibilidade para o *Big Data* é acompanhar indicadores estratégicos em tempo real. Quando a maioria dos dados era organizada e analisada manualmente, muitos aspectos

que podiam atrapalhar as vendas só eram descobertos mais tarde, quando já não era mais possível tomar providências a tempo de reverter à situação. Atualmente, os dados são processados a uma velocidade quase instantânea (FEIJÓ, 2013).

A principal razão da análise de *Big Data* é contribuir para que as empresas tomem melhores decisões de negócio. As análises de dados podem permitir uma abordagem de marketing direcionado possibilitando a empresa uma melhor compreensão de seus clientes, o que proporciona a vantagem competitiva que a maioria das empresas está buscando. (FEIJÓ, 2013).

Encontrar pessoas qualificadas para realizar a análise de dados é um dos principais desafios associados a *Big Data*. Iniciativas de sucesso nesta área envolvem a colaboração entre a Tecnologia da Informação (TI), os empresários e os cientistas de dados para identificar e implementar as análises que vão resolver os problemas de cada negócio. A ciência de dados é uma área de sucesso, e o cientista de dados é um novo tipo de profissional responsável por descobrir problemas complexos, ter a percepção e identificar as oportunidades de cada negócio. A demanda é alta por pessoas que podem ajudar a fazer sentido aos grandes fluxos de informação digital enviados pelas organizações (INTEL, 2012).

2.3 BIG DATA E O CONCEITO DOS 5 Vs

Existem até o presente momento, cinco conceitos que definem *Big Data*, também chamados de 5Vs:

a. Volume - O volume se refere à quantidade de informações digitais que são produzidas pelos usuários e/ou processos de aplicações a cada segundo. Esta é a característica que fica mais em evidência, pois o volume é a matéria-prima dessa nova tendência. (PETRY; VILICIC, 2013).
b. Variedade - Os dados que fazem parte desse imenso volume de informações, em sua grande maioria não possuem uma estruturação, são informações vindas de fontes mistas de dados, como, por exemplo, fotos, músicas, mensagens de celulares e eletrônicas, informações geoprocessadas, comentários em redes sociais, histórico de páginas visualizadas por meio de um navegador de Internet, cadeia de relacionamentos em uma rede social, ativações de leitoras de códigos de barras, entre outras (PETRY; VILICIC, 2013).

c. Velocidade – Este aspecto depende fortemente do grau de importância dos itens anteriores. Atualmente, a resposta em tempo real se tornou a grande necessidade e exigência de todos os envolvidos em um processo digital, citando como exemplo os setores público e privado, as áreas de telecomunicações, os trabalhadores e clientes, entre outros. (PETRY; VILICIC, 2013).
d. Veracidade - Segundo Zikipoulos (2012), *Big Data* pode ser caracterizada em quatro aspectos, sendo o quarto chamado de Veracidade. Este aspecto está relacionado ao fato de que os dados não são "perfeitos", no sentido de que é preciso considerar o quão bom devem ser os dados para que gerem informações úteis e também os custos para os tornar bons.
e. Validade - Alguns autores consideram um quinto aspecto, a validade dos dados, ou seja, sua vida útil, o tempo em que os mesmos precisam ser mantidos (TAUBE, 2012). Estes aspectos estão apresentados na Figura 1:

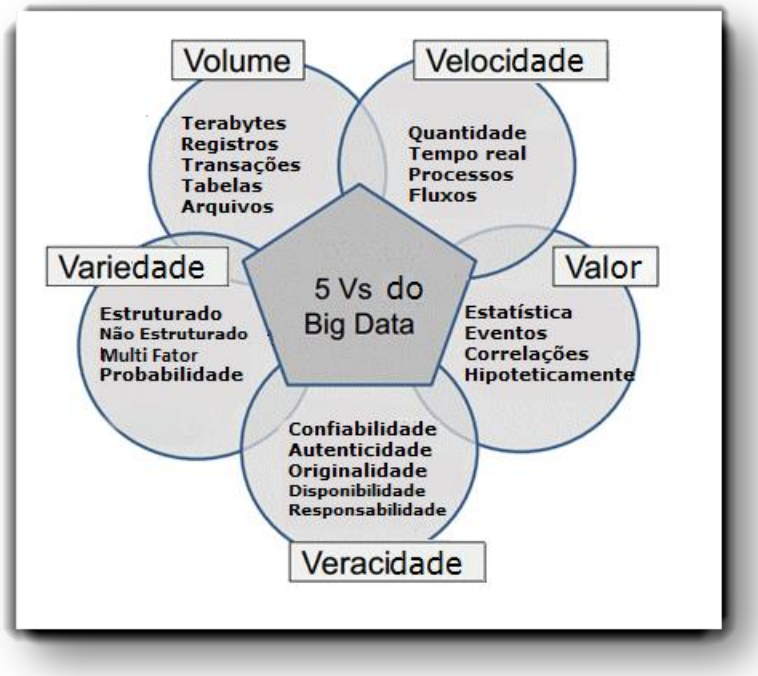


Figura 1 - *Big Data* e os 5Vs.
Fonte: World Newmedia Network (2013).

2.4 BIG DATA E DATA MINING

A mineração de dados (*Data Mining)* refere-se à atividade de analisar grandes conjuntos de dados para procurar informações relevantes ou pertinentes. As empresas coletam conjuntos enormes de dados que podem ser homogêneos ou coletados automaticamente. Os tomadores de decisão precisam ter acesso a partes menores e mais específicas a partir desses grandes conjuntos de dados (KAUSHIK PAL, 2014).

A mineração de dados é formada por um conjunto de ferramentas e técnicas que por meio do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatísticas são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. Esse conhecimento pode ser apresentado por essas ferramentas de diversas maneiras: agrupamentos, hipóteses, regras, árvores de decisão, grafos, entre outros. (KAUSHIK PAL, 2014).

Big Data é o ativo e mineração de dados é o "manipulador" desse ativo, que é utilizado para proporcionar resultados benéficos (KAUSHIK PAL, 2014).

A seguir na Figura 2 têm-se as principais características entre *Big Data* e *Data Mining*:

	Estatística	Data Mining	Big Data
Estrutura	estruturado	estruturado	desestruturado
Tamanho	pequeno	grande	muito grande
Geração	planejado	transacional	comportamental
Objetivo	compreender	otimizar os negócios	gerar negócios
Problemas de Privacidade	nenhum	menores	maiores
Fundada em	conceitos e teoria	tecnologia e ferramentas	tecnologia e ferramentas
Marketing	ruim	boa	perfeita

Figura 2 - Características *Big Data* x *Data Mining*
Fonte: *Big Data* Handler (2013).

2.5 DATA MINING E BUSISNESS INTELIGENCE

Diariamente as empresas acumulam grande volume de dados em seus aplicativos operacionais. São dados brutos que dizem quem comprou o quê, em que, quando e em que quantidade. É a informação vital para o dia-a-dia da empresa. Se for realizada a estatística ao final do dia para repor estoques e detectar tendências de compra, estará praticando-se *Business Intelligence* (BI). Se forem analisados os dados com estatística de modo mais refinado, à procura de padrões de vinculações entre as variáveis registradas, então estará sendo feita mineração de dados. (SHEPS, 2013).

Pode-se então diferenciar o *Business Intelligence* (BI) da mineração de dados (*Data Mining*) como dois patamares distintos de atuação. O primeiro busca subsidiar a empresa com conhecimento novo e útil acerca do seu meio ambiente e funciona no plano estratégico. O segundo visa obter a partir dos dados operativos brutos, informação útil para subsidiar a tomada de decisão nos escalões médios e altos da empresa e funciona no plano tático. (BERNARD, 2013).

2.6 BIG DATA E BUSINESS INTELLIGENCE

O termo *Business Intelligence* (BI) surgiu na década de 1980 no GartnerGroup e faz referência ao processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados contidos em *DataWarehouse/Data Mart*[10] gerando informações para o suporte à tomada de decisões no ambiente de negócios. (SHEPS, 2013).

Um *Data Warehouse* (DW) é simplesmente uma consolidação de dados a partir de uma variedade de fontes que se destina a apoiar a tomada de decisão estratégica e tática. Seu principal objetivo é fornecer uma visão coerente do negócio em um ponto no tempo. Usando vários conjuntos de ferramentas de armazenamento de dados, os usuários são capazes de executar consultas on-line e em seus dados. (SHEPS, 2013).

[10] Data mart (repositório de dados) é subconjunto de dados de um *Data warehouse* (armazém de dados).

Na Figura 3 tem-se o funcionamento do processo de BI:

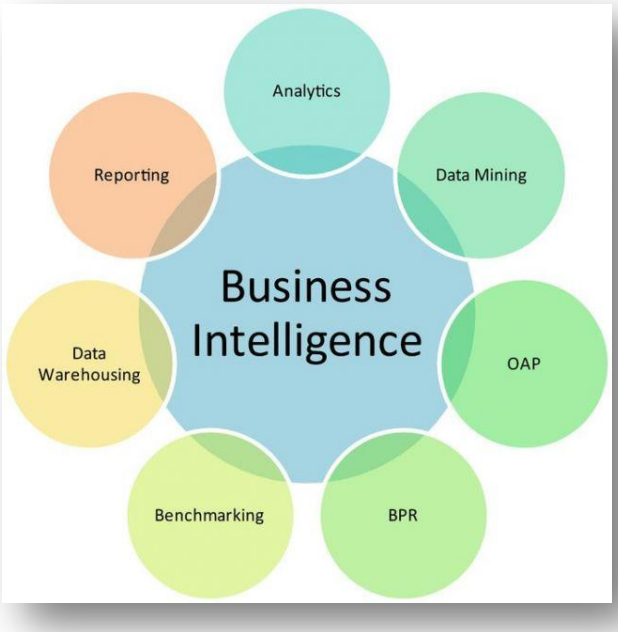


Figura 3 - Funcionamento do BI
Fonte: World Newmedia Network (2013).

- *Data Warehousing*: a integração de dados de uma ou mais fontes e assim, cria um repositório central de dados, um *Data Warehouse* ou armazéns de dados.
- *Data Mining*: o processo de explorar grandes quantidades de dados à procura de padrões consistentes para detectar relacionamentos e novos subconjuntos de dados a serem mapeados e extrair-se informações privilegiadas.
- *Analytics e Reporting*: as análises de minerações geram relatórios detalhados para fortalecer o esclarecimento do cenário.
- BPR: A Reengenharia de processos de negócio é uma estratégia de gestão de negócios para a análise e desenho dos fluxos de trabalho e dos processos de negócios que visa à reestruturação organizacional.
- *Benchmarking:* a busca das melhores práticas com o propósito de maximizar o desempenho.

A solução de BI tem foco na coleta, organização, transformação e disponibilização de dados estruturados para a tomada de decisão, além de permitir a análise preditiva de forma rápida e assertiva às organizações. Fornecem *insights* (ideias*)* e tendências aos gestores, para

assim poderem criar diretrizes eficazes para o alcance dos resultados empresariais almejados (ELIAS, 2014).

A implementação de uma solução de *Big Data* exige, uma boa maturidade em BI. Porque o *Big Data* possui grande complexidade e requer experiência em soluções que permitam uma fácil concepção no que diz respeito à análise de dados. Por isso, a implantação de *Big Data* sem uma experiência prévia em soluções analíticas aumenta, e muito, as chances de erro. (ELIAS, 2014).

2.7 TECNOLOGIAS E FERRAMENTAS ASSOCIADAS À BIG DATA

Nesta seção, são abordadas as principais tecnologias e ferramentas que estão sendo utilizadas nos últimos anos para apoio e desenvolvimento de *Big Data.*

2.7.1 Apache Hadoop

O *Hadoop* é um projeto que apresenta uma solução para problemas de *Big Data*. *Hadoop* é um framework para processamento distribuído e particionamento de grande volume de dados. (MASSUDA, 2013).

O *Hadoop* é um projeto *de Software* Livre, que foi desenvolvido pela *Apache Software Foundation*[11]e, portanto, permite a criação de um ecossistema de negócios baseados em distribuições específicas. O surgimento de serviços em nuvem, como o *AmazonElasticMapReduce*[12], permite às empresas tratarem dados massivos sem demandar aquisição de servidores físicos. Neste modelo, o usuário escreve a aplicação *Hadoop* e roda em cima da nuvem da Amazon[13]. (TAURION, 2012).

Hadoop é um *framework* de código aberto, implementado em Java e utilizado para o processamento e armazenamento em larga escala, para alta demanda de dados, utilizando

[11] Apache *Software* Foundation é uma plataforma de computação distribuída, com alta escalabilidade, grande confiabilidade e tolerância a falhas. (HTTP://apache.org).

[12] O Amazon Elastic MapReduce (Amazon EMR) é um serviço da *web* que permite processar vastas quantidades de dados facilmente, rapidamente e com um bom custo benefício.

[13] Amazon.com é uma empresa multinacional de comércio eletrônico dos Estados Unidos com sede em Seattle, estado de Washington. Foi uma das primeiras companhias com alguma relevância a vender produtos na Internet.

máquinas comuns. *Hadoop* na prática é uma combinação de dois projetos separados: o *HadoopMapReduce* , que é um *framework* para processamento paralelo, e o *HadoopDistributed File System* (HDFS).(TAURION, 2012),conforme abaixo na Figura 4:

Figura 4 - Arquitetura Hadoop.
Fonte: HortonWorks, Inc (2013).

O *MapReduce* é um *framework* computacional para processamento paralelo criado pelo Google. Ele abstrai as dificuldades do trabalho com dados distribuídos, trazendo referências da arquitetura *shared-nothing,* que é uma computação distribuída na qual cada nó é independente e autossuficiente, e não há um único ponto de discórdia em todo o sistema, eliminando quaisquer problemas que poderiam ocorrer com o compartilhamento de informações. (MASSUDA. 2013).

Segundo Gasparotto (2013), HDFS é um sistema de arquivos distribuído criado para armazenar arquivos muito grandes. O conceito sobre o qual o HDFS foi construído é chamado de escreva uma vez (*write-once*) e leia muitas vezes (*read-many-times*). É ideal para o *Hadoop*, onde seus os dados são escritos apenas uma vez, mas são processados diversas vezes, dependendo da aplicação.

2.7.2 Jasper soft BI Suite

O pacote *Jaspersoft* é um dos líderes de código aberto para a produção de relatórios de colunas de banco de dados. O *software* foi instalado em muitas empresas em que geram tabelas SQL em PDFs que todos possam escrutinar nas reuniões.

O *JasperReports* Server oferece *software* para obter dados de muitas das principais plataformas de armazenamento, incluindo MongoDB[14],*Cassandra*[15], *Redis*[16], *Riak*[17], *CouchDB*[18] , *e Neo4j*[19]. *Hadoop* também está bem representado, com *JasperReports* que fornece um conector *Hive*[20] para chegar dentro de *HBase*. (JASPERSOFT, 2014).

2.7.3 Pentaho Business Analytics

Trata-se de um conjunto de ferramentas criadas em cima do conceito de inteligência de negócios (Business Intelligence – BI). Essas ferramentas permitem realizar:

- Geração de relatórios empresariais;
- ETL (Extração, Transformação e Carga);
- Análise de informações (OLAP);
- Painéis de controle (Dashboards);
- Mineração de dados (*Data Mining*)

A plataforma Pentaho, conhecida como Pentaho Open BI, é composta de aplicações open source para criação de soluções de Business Intelligence. (OLIVEIRA, 2015).

[14] Disponível em: www.mongodb.org/
[15] Disponível em: www.cassandra.apache.org/
[16] Disponível em: http://redis.io/
[17] Disponível em: http://basho.com/riak/
[18] Disponível em: http://couchdb.apache.org/
[19] Disponível em: http://neo4j.com/
[20] Disponível em: https://hive.apache.org/

Entre os módulos destacam-se (Figura 5):

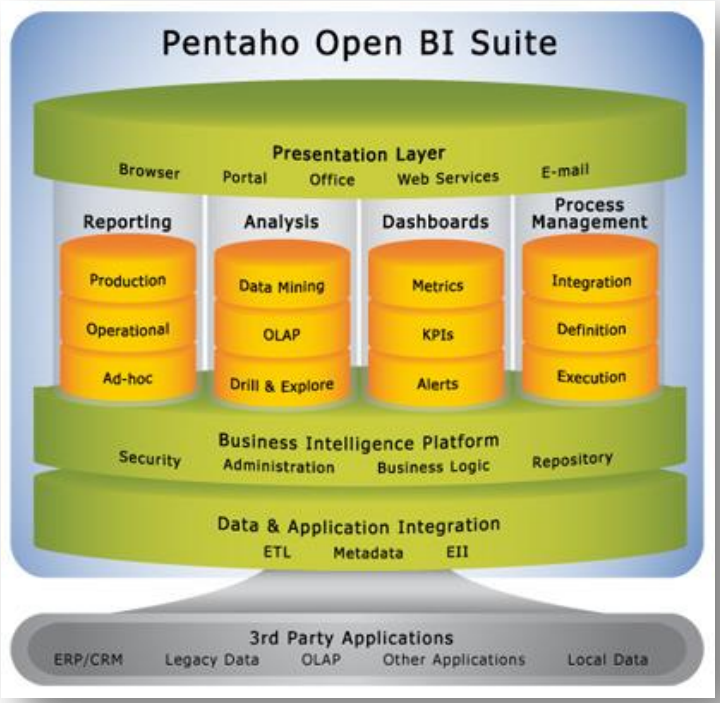


Figura 5 - Estrutura Pentaho.
Fonte: Pentaho (2015).

- *BI Server:* Front-end de interação com o usuário final. Provê dois "auto-serviços" conhecidos como PUC *(Pentaho User Console)* e PAC *(Pentaho Administration User*). O WAQR (*Web AdHoc Query and Reporting*) permite criação de relatórios on-line, via *web*, e o OLAP *(OnLine Analytical Processing)* permite navegação por meio de níveis (nível agregado ao menor grão). E em se tratando de BI, o OLAP é o coração da análise e dá acesso aos relatórios publicados pelo *Report Designer*.
- *Report Designer:* Esta ferramenta deve ser utilizada para criação de relatórios complexos. Quando for necessário criar relatórios mais interativos e elaborados esta ferramenta é a mais adequada, pois fornece mais recursos que a ferramenta de geração de relatórios ad-hoc.
- *Design Studio*: Esta é uma ferramenta que é baseada no Eclipse (API para desenvolvimento). Um ambiente de desenvolvimento de soluções avançadas de BI.

- *Aggregation Designer*: Uma ferramenta gráfica que ajuda a melhorar a eficiência do cubo *Mondrian*, criando tabelas agregadas. Porém seu uso deve ter cuidados.
- *Metadata Editor:* Ferramenta que mapeia os bancos de dados e seu conteúdo em uma visão de negócios a partir da qual usuários podem compor relatórios, via *web*, por meio do WAQR. Permite adição de uma camada de *metadados* a uma fonte de dados existente. Normalmente usada para produzir uma camada que facilita a criação de relatórios ou análises, porém seu uso não é obrigatório.
- *Pentaho Data Integration*: Ferramenta ETL (antigo projeto Kettle), que permite acessar e preparar fontes de dados para análise, mineração e geração de relatórios OLAP. Ele é normalmente iniciado quando se quer preparar dados para análise.
- *Pentaho Schema Workbench*: Uma ferramenta gráfica que realiza a criação de esquemas ROLAP para análise. Este é um passo necessário para preparar os cubos. Ele possui integração com BI Server e permite publicar o esquema desenvolvido diretamente nele. Tem a licença gratuita e a opção paga com mais recursos. Apresentam indicadores de acesso *web*, independentes do SO que estará sendo utilizado.
- Personalização nos indicadores de negócio *(Dashboard*, cubos e relatórios).

2.7.4 Karmasphere Studio

Karmasphere Studio é um conjunto de plug-ins construídos em cima de Eclipse. É uma IDE especializada para criar e executar tarefas do *Hadoop.*

Para configurar uma tarefa *Hadoop* com esta ferramenta, há uma série de etapas em um trabalho *Hadoop* e ferramentas de *Karmasphere* para orientá-lo por meio de cada passo, que mostra os resultados parciais ao longo do caminho, como configurar o fluxo de trabalho, as ferramentas que exibem o estado dos dados de teste em cada etapa. Os dados temporários serão cortados, analisados e, em seguida, reduzidos. (KARMASPHERE INC, 2014).

2.7.5 Talend Open Studio

Talend também oferece um *IDE* baseado em Eclipse para trabalhos de processamento de dados com Hadoop. Suas ferramentas são projetadas para ajudar com a integração, qualidade e gerenciamento de dados, todos com sub-rotinas ajustadas para estes postos de trabalho *(TALEND*,2014).

2.7.6 Skytree Servidor

Skytree Server executa uma série de clássicos algoritmos de aprendizado em seus dados, usando este servidor a empresa pode ser 10.000 vezes mais rápido do que outros pacotes podem pesquisar os dados em busca de grupos de itens matematicamente semelhantes, em seguida, inverter isso para identificar valores discrepantes que podem ser problemas, oportunidades ou ambos. Os algoritmos podem ser mais precisos do que os humanos, e eles podem pesquisar grandes quantidades de dados em busca de entradas que são um pouco fora do comum. (SKYTREE INC, 2014).

.

2.7.7 Tableau Software

Tableau Software começou a utilizar *Hadoop* várias versões atrás, Tableau depende do *Hive* para estruturar as consultas, em seguida, tenta armazenar em cachê o máximo de informações na memória para permitir que a ferramenta seja interativa. Embora muitas das outras ferramentas de relatórios sejam construídas sobre uma tradição de gerar os relatórios off-line, Tableau oferece um mecanismo interativo para que se possam dividir seus dados novamente. *(TABLE SOFTWARE, 2014).*

2.7.8 Splunk

Splunk é um pouco diferente das outras opções. Não é exatamente uma ferramenta de geração de relatório ou um conjunto de rotinas de IA, ele cria um índice dos seus dados, como se os dados fossem um livro ou um bloco de texto. *(SPLUNK INC, 2014).*

2.8 CLOUD COMPUTING E BIG DATA

De acordo com Ambrust et al. (2009), *Cloud Computing* refere-se a aplicações entregues como serviços por meio da Internet, *hardwares e softwares* de sistemas nos *data centers* que prestam esses serviços.

Cloud Computing ou computação em nuvem, oferece um ambiente para pequenas e médias empresas para implementar a tecnologia *Big Data*. Benefícios que as empresas podem realizar a partir de grandes volumes de dados incluem a melhoria do desempenho, fazendo de apoio à decisão e inovação em modelos de negócios, produtos e serviços (MANYIKA et al., 2011). A Figura 6 demonstra uma visão geral da Computação em Nuvem.

Figura 6 - Computação em Nuvem
Fonte: Wikipedia (2013).

3 MATERIAL E MÉTODOS

Este capítulo apresenta uma descrição das principais ferramentas utilizadas para a análise de *Big Data*.

3.1 MAPREDUCE

MapReduce é um modelo de programação que processa grandes conjuntos de dados. Foi construído um sistema em torno deste modelo de programação em 2003, simplificando a construção dos índices para lidar com buscas no *Google.com*. Desde então, mais de 10.000 programas distintos foram implementados utilizando o *MapReduce* do *Google*, incluindo algoritmos de grande escala de processamento gráfico, processamento de texto, *machine learning*[21], e tradução estatística automática. O *Hadoop* é uma implementação *open source* de *MapReduce* e tem sido amplamente utilizado fora do *Google* por uma série de organizações (DEAN; GHEMAWAT, 2010).

A capacidade de armazenamento dos elementos aumentou muito nos últimos anos, entretanto, a velocidade de leitura e escrita não seguiu este ritmo. Segundo Gasparotto (2013), existem dois tipos de problemas. O primeiro é que se existirem 100 vezes mais discos rígidos, a chance de existir falha em um deles é 100 vezes maior o que pode trazer a perda de dados. Uma maneira de evitar esse tipo de problema seria utilizando-se a replicação, por meio de cópias de segurança dos dados que são mantidas em diferentes discos. O segundo problema é que muitas tarefas de análise de dados precisam combinar dados "espalhados" em discos diferentes. Todavia, o *MapReduce* resolve este problema, oferecendo um modelo de programação que abstrai os dados, uma vez que o processamento é realizado por meio de uma combinação entre chaves e valores (*keys* e *values*) que podem estar em diferentes discos rígidos.

[21] **Machine learning**: capacidade da máquina de reconhecer padrões que ocorreram várias vezes e melhorar seu desempenho com base em experiências passadas.

Segundo Xiao (2013), *MapReduce* consiste das seguintes funções (Figura 7):

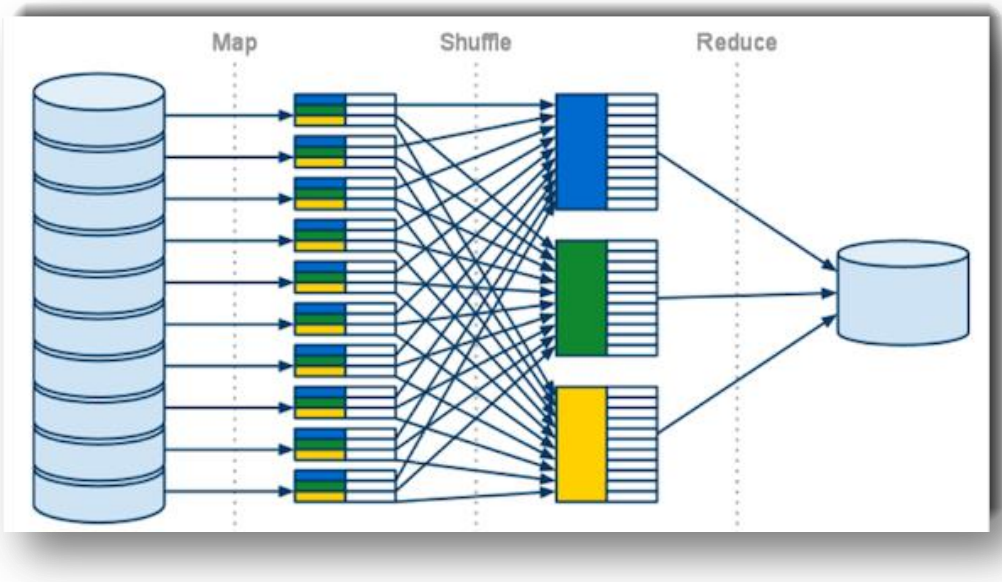


Figura 7 - Arquitetura MapReduce
Fonte: Artificial Intelligence in Motion (2012).

- *Map*: Responsável por receber os dados na forma de *key/value* representando de forma lógica cada registro dos dados de entrada, podendo ser, por exemplo, uma linha em um arquivo de *log* ou de uma tabela. A função *map* retorna uma lista com zero ou mais dados chave/valor e deve ser codificada pelo desenvolvedor, por meio de outras ferramentas ou da API Java;
- *Shuffle:* A etapa de *shuffle* é responsável por organizar o retorno da função *Map*, atribuindo para a entrada de cada *Reduce* todos os valores associados a uma mesma chave. Esta etapa é realizada pela biblioteca do *MapReduce*;
- *Reduce:* Por fim, ao receber os dados de entrada, a função *Reduce* retorna uma lista de chave/valor contendo zero ou mais registros, semelhante ao *Map*, deve ser codificada pelo desenvolvedor.

A arquitetura do *MapReduc*e é semelhante ao do HDFS (*Hadoop Distributed File System*), que é um sistema de arquivos criado para armazenar arquivos muito grandes de forma distribuída. No *MapReduce* os componentes são:

- *JobTracker:* Ele recebe o *job MapReduce* e programa as tarefas *map* e *reduce* para execução, coordenando as atividades nos *TaskTrackers;*

- *TaskTracker:* Componente responsável por executar as tarefas de *map* e *reduce* e informar o progresso das atividades.

O *HadoopMapReduce* é o responsável pelo processamento de dados, podendo executar várias tarefas, fazendo um envio de um arquivo de entrada para um arquivo de saída. Como exemplo, tem-se a contagem de palavras em documentos diferentes. Sem o uso do *MapReduce*, o programador teria alguns problemas que são bem comuns no processamento de dados. Mas com o uso de *MapReduce*, o *Hadoop* resolve para o desenvolvedor, evitando assim problemas de escalonamento. (GASPAROTTO, 2013).

Segundo Goldman (2013), atualmente, muitas empresas já utilizam o *Hadoop* para diversos fins, seguem algumas delas:

- *Adobe* (www.adobe.com): possui ferramentas e serviços para conteúdo digital, usa *Hadoop* no armazenamento e processamento de dados internos e de redes sociais;
- *E-Bay* (www.ebay.com): é um comércio eletrônico com foco em uma plataforma global de negociação (shopping popular); também usa na otimização de buscas;
- *Facebook* (www.facebook.com): rede social utiliza *Hadoop* para análise de log;
- *Last. FM* (www.last.fm): rádio online que utiliza *Hadoop* para análise de log e análise de perfil de usuário;
- *LinkedIn* (www.linkedin.com): é uma rede social que compartilha informações, ideias e oportunidades. Em que usa: análise e busca de similaridade entre perfis de usuários;
- *The New York Times* (www.nytimes.com): uma das maiores empresas jornalísticas mundiais. Utiliza *Hadoop* para conversão de imagens, armazenamento de jornais digitais;
- *Twitter* (www.twitter.com): é uma rede social, que usa no armazenamento de mensagens e no processamento de informações;
- *Yahoo!* (www.yahoo.com): oferecem serviços de busca na *Web*, serviços de notícias e e-mail. Em que usa: no processamento de buscas, recomendações de publicidades, testes de escalabilidade.

3.2 HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

O HDFS possui as características básicas existentes em um sistema de arquivos. Sendo assim, os arquivos possuem permissões para escrita, leitura e execução. Além disso, possui um log de edição de dados, para que se tenha um controle maior nas modificações que ocorrem nos dados do sistema de arquivos (GASPAROTTO, 2013).

O HDFS precisa de dois tipos de nós de armazenamento para seu funcionamento: um mestre *(namenode*) e um ou mais trabalhadores (*datanodes*). O *namenode* lidera todo o sistema de arquivos, onde mantem metadados para todos os arquivos e arquivos do sistema. O mestre sabe também quais *datanode*s os blocos de um determinado arquivo possuem. Os trabalhadores possuem a função de guardar e recuperar blocos, quando mandados pelo *namenode*, e enviar relatórios para ele periodicamente, com as listas dos blocos que os *datanodes* estão armazenando. (GASPAROTTO, 2013). A arquitetura do HDFS pode ser vista na Figura 8 abaixo:

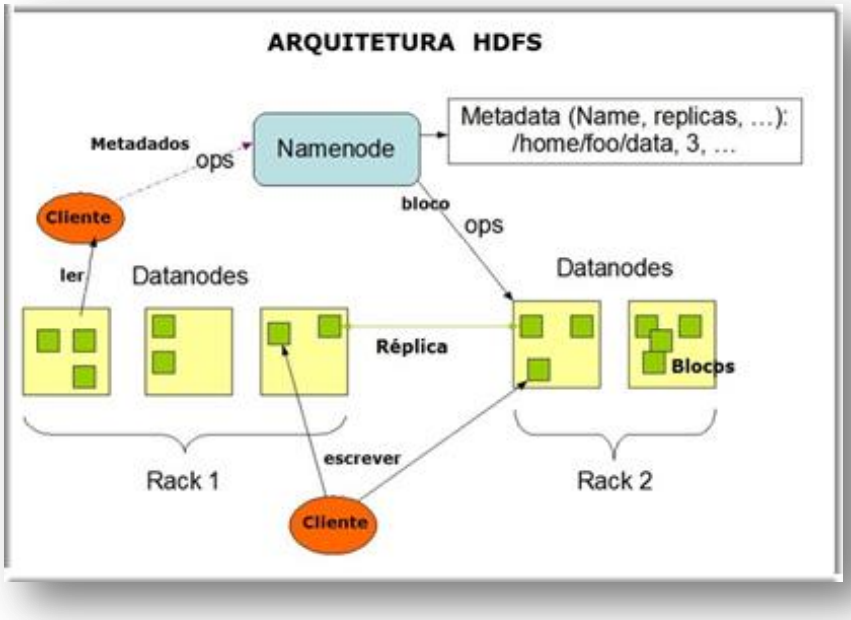


Figura 8 - Funcionamento do HDFS
Fonte: Blog SpecIndia (2013).

Segundo *Apache Foundation*, 2013, os subprojetos mais conhecidos do *Hadoop* são o *MapReduce* e o HDFS, mas além deles existem outros que oferecem uma série de serviços extras, facilitando assim o desenvolvimento. Entre eles tem-se:

- *Avro*: é um sistema de serialização de dados que fornece uma estrutura de dados rica, um formato compacto, rápido e binário de dados, um arquivo recipiente para armazenar dados persistentes, chamada de procedimento remoto (RCP), integração simples com linguagens dinâmicas;
- *Pig*: uma plataforma para análise de grandes conjuntos de dados que consiste em uma linguagem de alto nível para expressar programas de análise de dados. Além disso, a sua camada de infraestrutura consiste em um compilador que produz sequencias de programas *MapReduce;*
- *HBase*: uma base de dados distribuída, criada para armazenar tabelas muito grandes. Trata-se de um modelo de armazenamento orientado a colunas, muito fácil de utilizar com a API Java;
- *ZooKeeper*: é um serviço centralizado para manter as informações de configuração das aplicações, proporcionando configuração distribuída e fornece uma interface simples para auxiliar o desenvolvedor, evitando bugs e condições de corrida (*race conditions*);
- *Hive*: *software* de *Data Warehouse* distribuído, facilita a consulta e gerenciamento de grandes conjuntos de dados (*datasets*) em ambientes de armazenamento paralelo. Fornece uma linguagem baseada em SQL, chamada *HiveQL*, que serve para facilitar a estruturação dos dados e pesquisa. A figura 9 apresenta estas características:

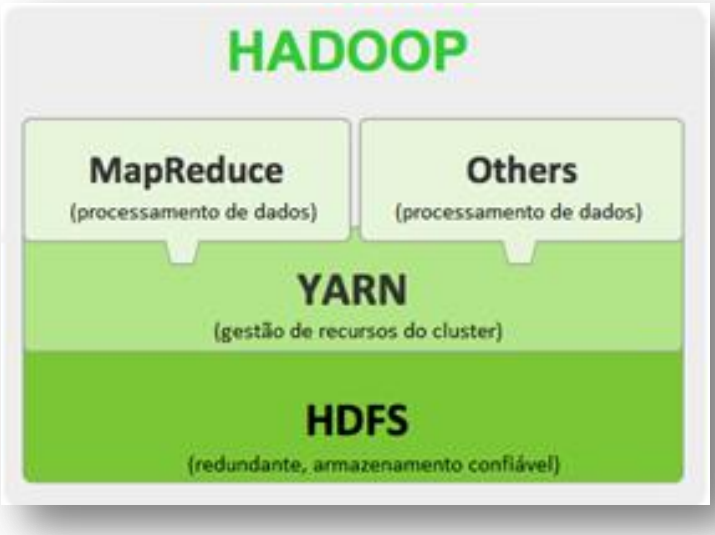


Figura 9 - Hadoop e suas Características
Fonte: *Big Data* Handler (2013).

3.3 HORTONWORKS DATA PLATAFORM

A plataforma de dados *Hortonworks* (HDP) é uma distribuição de código aberto, distribuído por *Apache Hadoop*. O HDP fornece as versões dos componentes do Apache, liberadas, com todas as correções de bugs necessárias para tornar todos os componentes interoperáveis em seus ambientes de produção. (Figura 10):

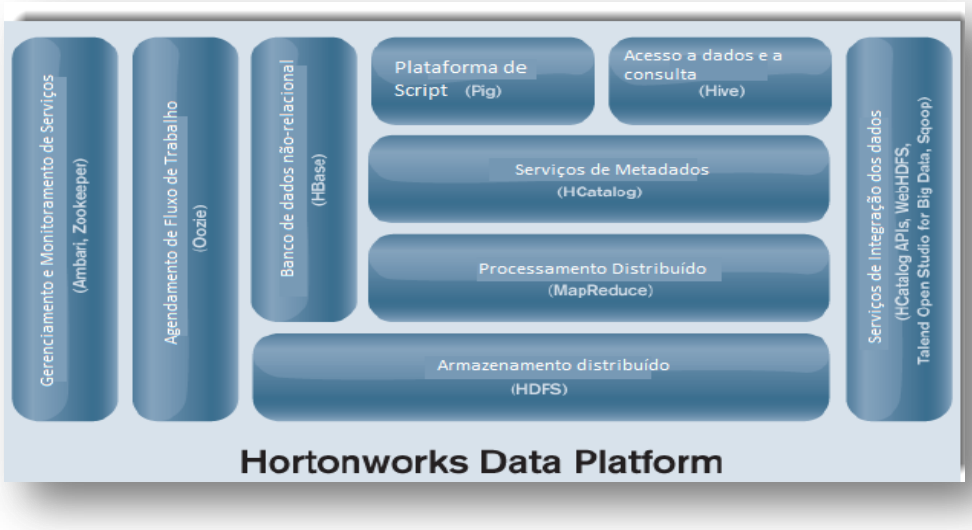


Figura 10 - Hortonsworks Data Platform
Fonte: Hortonworks, Inc (2013).

- Plataforma central Hadoop (HDFS, MapReduce);
- Banco de dados não relacional (Apache HBase);
- Serviços de metadados (Apache HCatalog);
- Plataforma de script (Apache Pig);
- Acesso a dados e consulta (Apache Hive);
- Agendador de Fluxo de Trabalho (Apache Oozie);
- Coordenação de Cluster (Apache Zookeeper);
- Gestão e monitorização (Apache Ambari);
- Serviços de integração de dados (HCatalog APIs, WebHDFS, Talend Open Studio para Big Data, e Apache Sqoop);
- Serviços de gerenciamento de log distribuídos (Apache Flume);
- Biblioteca de aprendizado de máquina (Mahout);

3.3.1 Hortonworks Sandbox

O *Sandbox* é um cluster *single node Hadoop* executado em uma máquina virtual, é um ambiente *Hadoop* pessoal, que inclui desenvolvimentos da última distribuição HDP, reunidos em um ambiente virtual.

É necessária a utilização da máquina virtual *VirtualBox*, para o funcionamento da *Sandbox* (Figura 11):

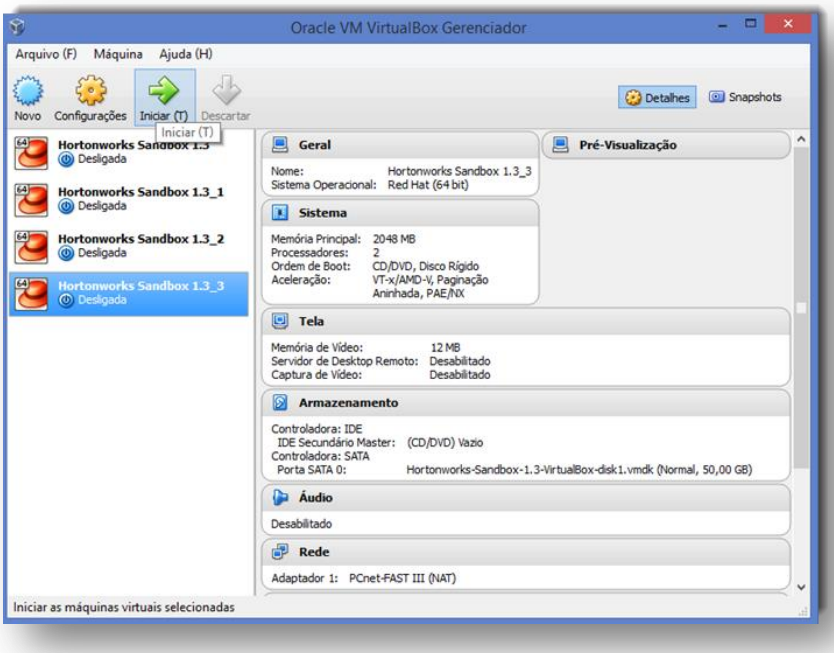


Figura 11 - Iniciando a *Hortonworks Sandbox* na *Virtual Box*.
Fonte: Autoria própria.

O console mostra as instruções de *login* para a *Sandbox*. (Figura 12).

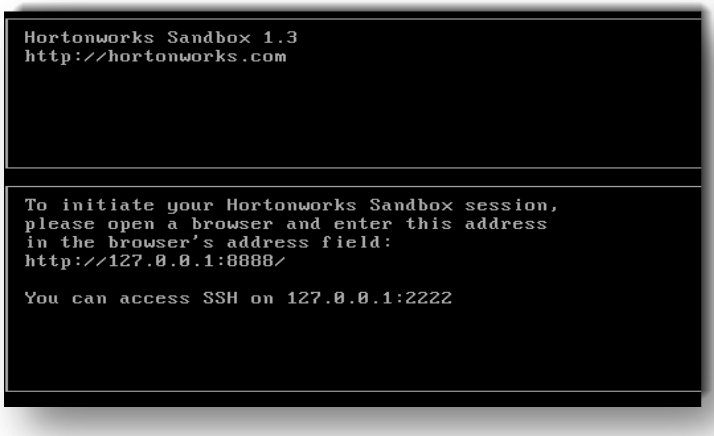


Figura 12 - *Console Sandbox*
Fonte: Autoria própria.

No *browser* deve-se digitar o endereço que aparece no console *Sandbox*, em seguida abrirá a página inicial (Figura 13).

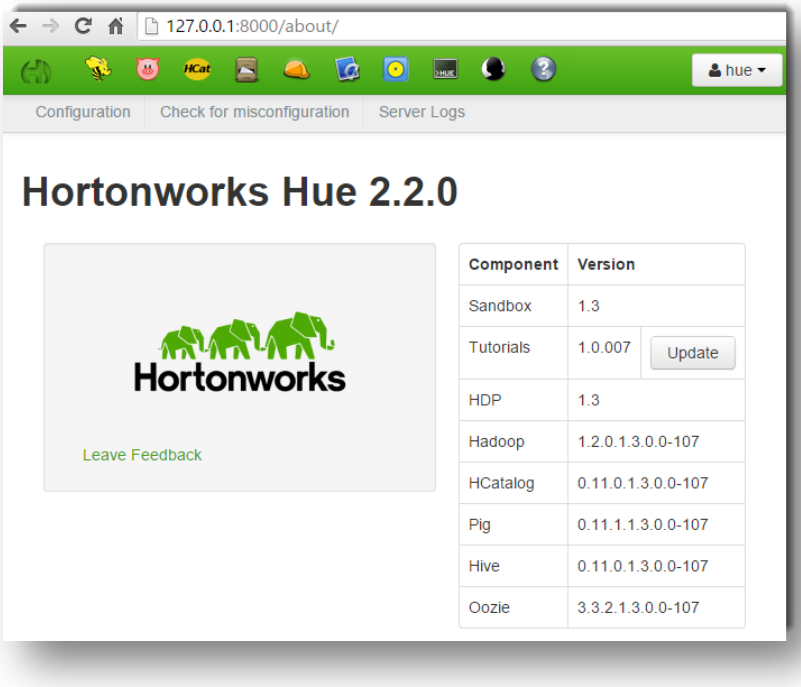


Figura 13 - Página Inicial Sandbox
Fonte: HortonWorks (2013).

3.4 ESCOPO DO ESTUDO

Dados de sentimento são dados não estruturados que representam opiniões, emoções e atitudes contidas em fontes, tais como mensagens de mídia social, blogs, análises de produtos on-line, e interações de suporte ao cliente.

Neste estudo, são utilizados os dados de sentimento referentes à opinião dos usuários sobre o filme *Iron Man* 3 (O Homem de Ferro 3) um filme de ação do ano de 2013, baseado no personagem fictício Homem de Ferro produzido pela Marvel Studios.com a finalidade de analisar como o público se sentiu em relação a este filme, se este produto teve um *feedback* positivo ou negativo entre seus usuários. Os dados obtidos referem-se a coleta efetuada junto ao ano de lançamento.

Com base na Figura 14 têm-se os dados de sentimentos referentes aos comentários postados via *Twitter* pelo público que assistiu ao filme. Este arquivo contendo as informações brutas quanto aos dados coletados é disponibilizado pela Hortonworks Sandbox.

É importante frisar que a partir deste arquivo disponibilizado, é feita toda a análise relativa a identificação de sentimento, conforme descrito nas seções a seguir.

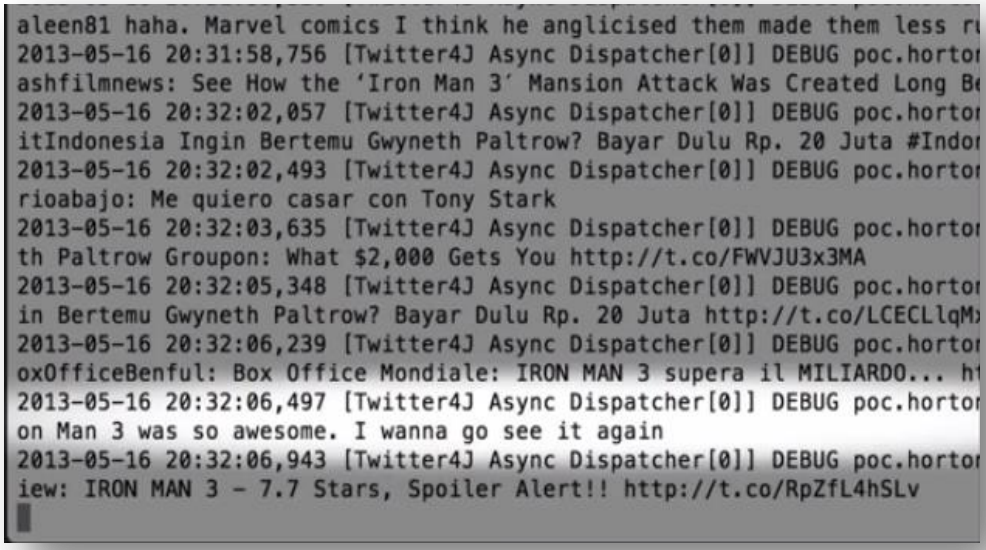


Figura 14- Amostra de dados de sentimento via Twitter
Fonte: HortonWorks (2013).

3.4.1 Obter os dados de sentimento por meio do Apache Flume

Após a obtenção do arquivo bruto, é preciso acessar a Sandbox para obter os dados de sentimento do *Twitter*. Neste estudo, esses dados foram obtidos por meio do *Apache Flume*, um serviço que pode ser usado para coletar, agregar e movimentar grandes quantidades de dados de *streaming* para o *Hadoop Distributed File System* (HDFS). O Flume é formado pelos seguintes componentes:

- Event: Unidade de dados que é transportada pelo *Flume*;
- Source: Entidade onde estão os dados coletados por *Flume;*
- Sink: É a entidade onde os dados são entregues a um destino;
- Channel: é o canal entre o Source e o Sink;
- Agent: qualquer máquina virtual onde o *Flume* for executado;
- Client: é a entidade que produz e transmite o evento para o Source, operando com o agente.

Os componentes do Flume interagem na seguinte maneira:

- Um fluxo em *Flume* começa a partir do *Client;*
- O *Client* transmite o evento a um *Source* operando dentro do *Agent*;
- O *Source* recebe este *Event*, em seguida, entrega para um ou mais *Channels;*

Na Figura 15 tem-se definidas as Sources, Sinks, Channels, e o *Agent*, neste caso o Agente é o "TwitterAgent", as "Sources" ou fonte de dados são Twitter 1 e 2, " Sink" onde os dados serão entregues será o HDFS. As chaves de acesso "*consumerKey", "consumerSecret", "acessToken" e "acessTokenSecret*", são geradas pela API do *Twitter* para que se possam ser extraídos os *tweets* corretamente.

```
witterAgent.sources = Twitter1, Twitter2
witterAgent.channels = MemChannel
witterAgent.sinks = HDFS

witterAgent.sources.Twitter1.type = poc.hortonworks.flume.source.twitter.TwitterSource
witterAgent.sources.Twitter1.channels = MemChannel
witterAgent.sources.Twitter1.consumerKey = TT1e0jpRS3Apb3G63KPBw
witterAgent.sources.Twitter1.consumerSecret = IAi3HtCsXUs61AZbtdjaOsVLUazx5D2tYqSZsxhndzk
witterAgent.sources.Twitter1.accessToken = 16566317-MXoIpwwcyI2PDAvISwnjeb2SNB7RBQeDuyZKytmbS
witterAgent.sources.Twitter1.accessTokenSecret = F9m9H6O462JXv2isSX10gA88rLVW7jGwOduYQi6mU
witterAgent.sources.Twitter1.keywords = Hortonworks, Hadoop, Big Data, Owen O'Malley, Arun Murthy,
Herb Cunitz, open source, Apache Software Foundation, Cloudera, Impala, Hive, Stinger, MapR
```

Figura 15 - Configuração Twitter
Fonte: Hortonworks (2013).

Na Figura 16 tem-se a instalação do Flume na Sandbox:

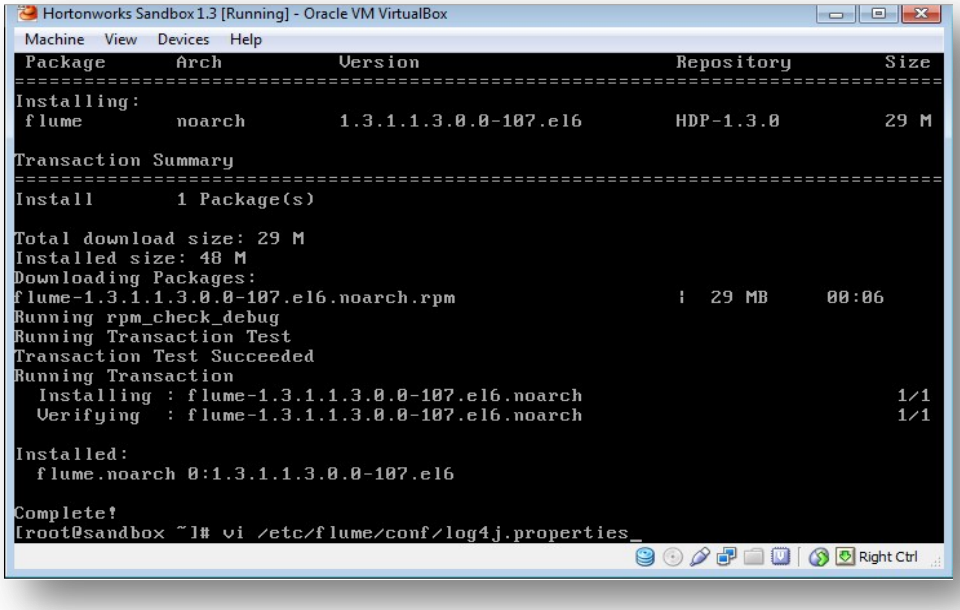


Figura 16- Instalando o Flume na Sandbox
Fonte: Autoria Própria.

Para iniciar o *Flume* deve-se utilizar o seguinte comando: `flume-ng agent -c /etc/flume/conf -f /etc/flume/conf/flume.conf -n sandbox` (Figura 17).

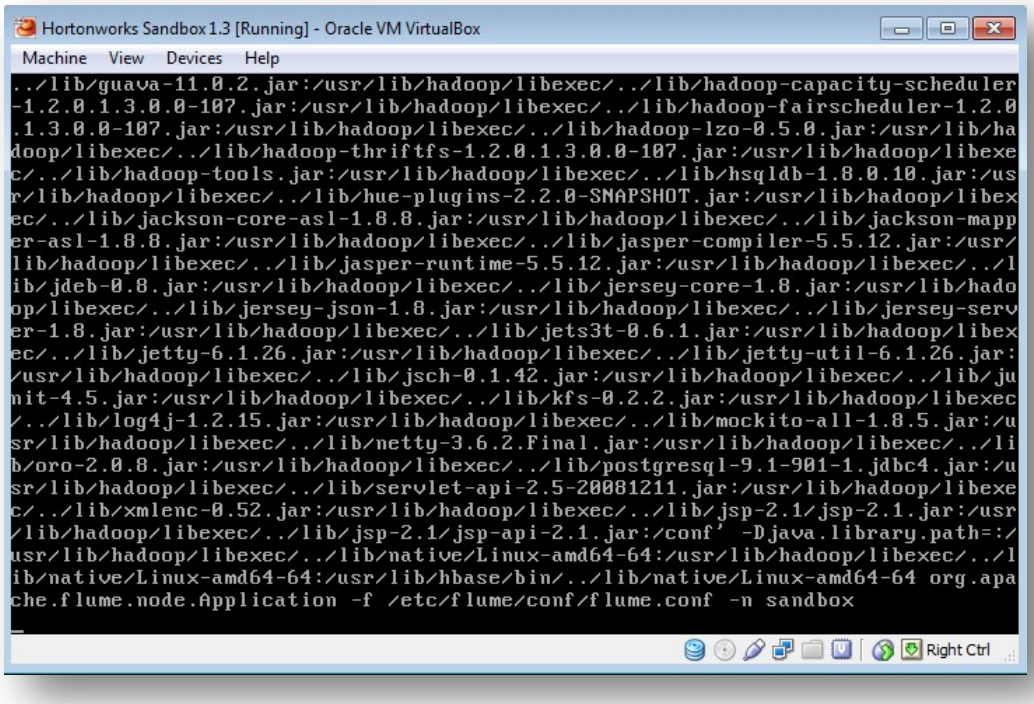


Figura 17 - Flume em execução
Fonte: Autoria Própria.

Com os dados já extraídos pelo *Flume*, é possível carregá-los na *Sandbox*. Tem-se um arquivo.zip com todos os arquivos de dados necessários para o *upload (*Figura 18):

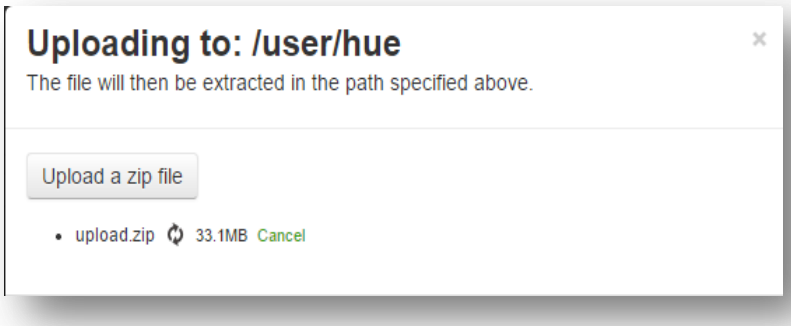


Figura 18 - Carregamento do arquivo .zip upload
Fonte: Autoria Própria.

Quando o *upload* estiver completo, o arquivo poderá ser visto no diretório *Sandbox (*Figura 19):

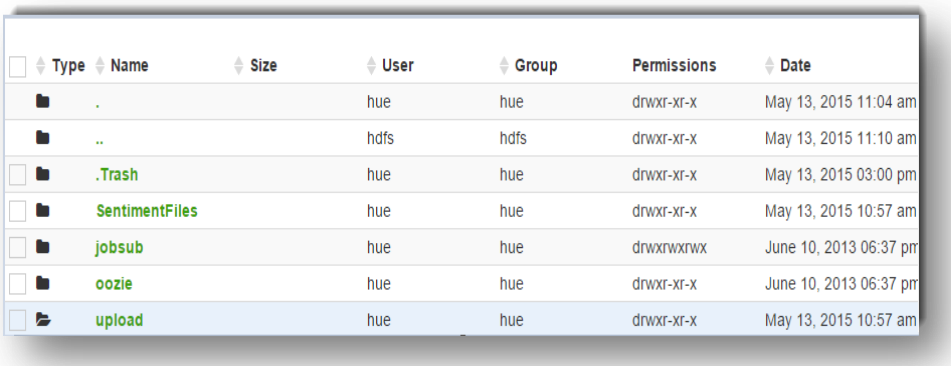


Figura 19 - Arquivo extraído no diretório
Fonte: Autoria Própria.

É necessário transferir um script *Hive* para o diretório *Sandbox*, para isso utiliza-se o aplicativo WinSCP (Windows Secure CoPy) é um SFTP livre e de código aberto, para Microsoft Windows. Sua principal função é a transferência segura de arquivos entre um local e um remoto (Figura 20):

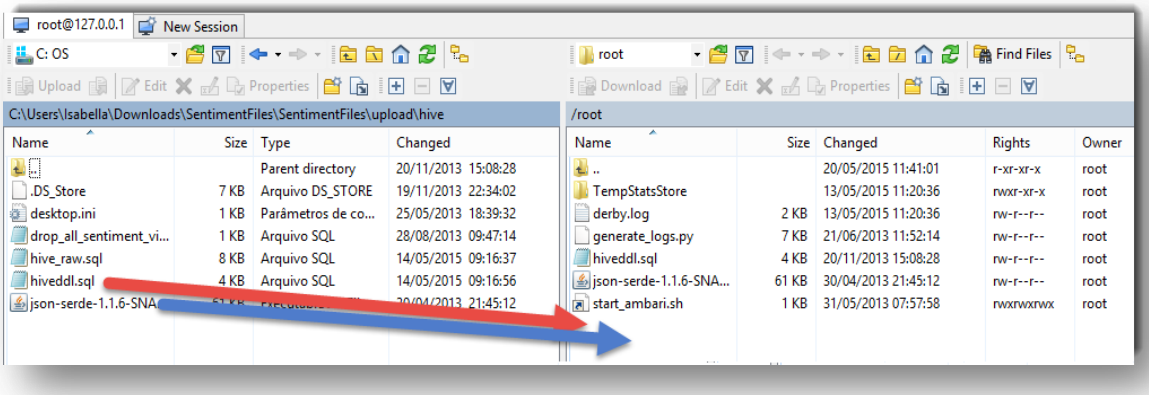


Figura 20 - Transferência de arquivos com WinSCP
Fonte: Autoria Própria.

3.4.2 Refinando os dados com MapReduce

Para refinar os dados brutos extraídos do *Twitter*, é necessário executar o *Script Hive* que foi transferido para o diretório *Sandbox* anteriormente. Precisam-se executar os seguintes comandos: `Alt + F5`, em seguida, efetuar login no Sandbox usando login: `root` e Senha:

`hadoop`, o *prompt* de comando aparecerá com o prefixo `[root @ sandbox ~] #`: (Figura 21).

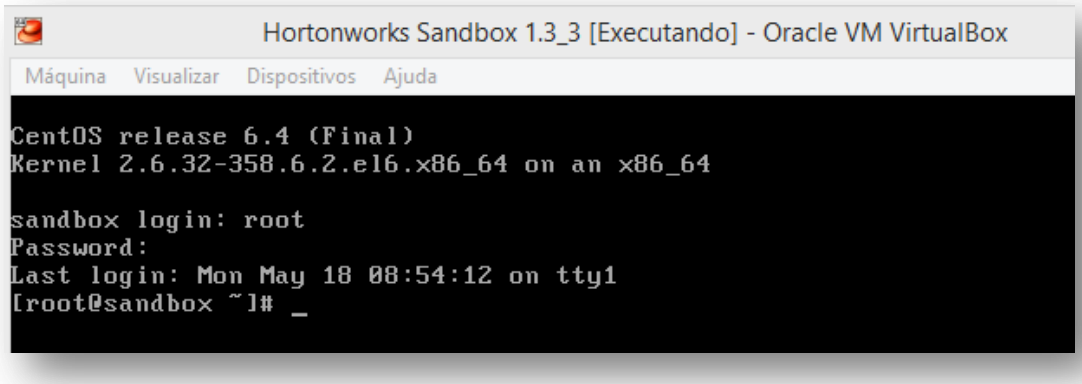


Figura 21 - Login no Console Sandbox
Fonte: Autoria Própria.

Deve se digitar em seguida o comando: `hive-f hiveddl.sql.`

Este *script* tem a função de executar uma série de trabalhos *MapReduce* que é responsável por refinar os dados brutos. (Figura 22)

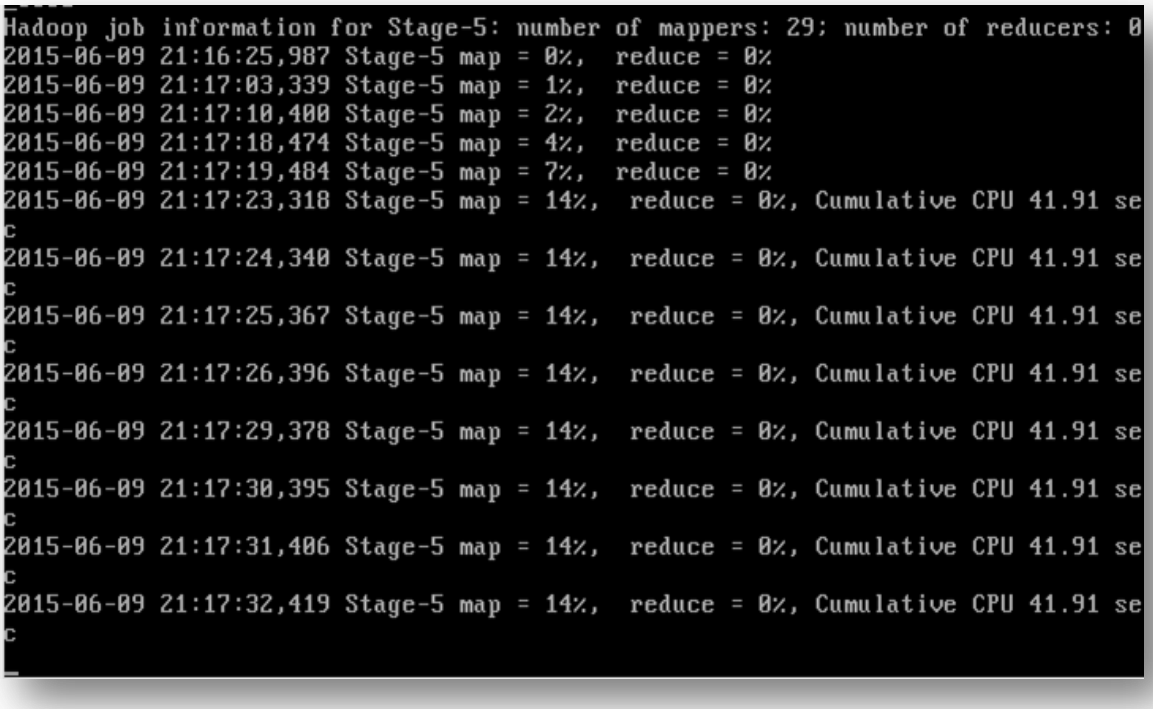


Figura 22 - Execução do script Hive
Fonte: Autoria Própria.

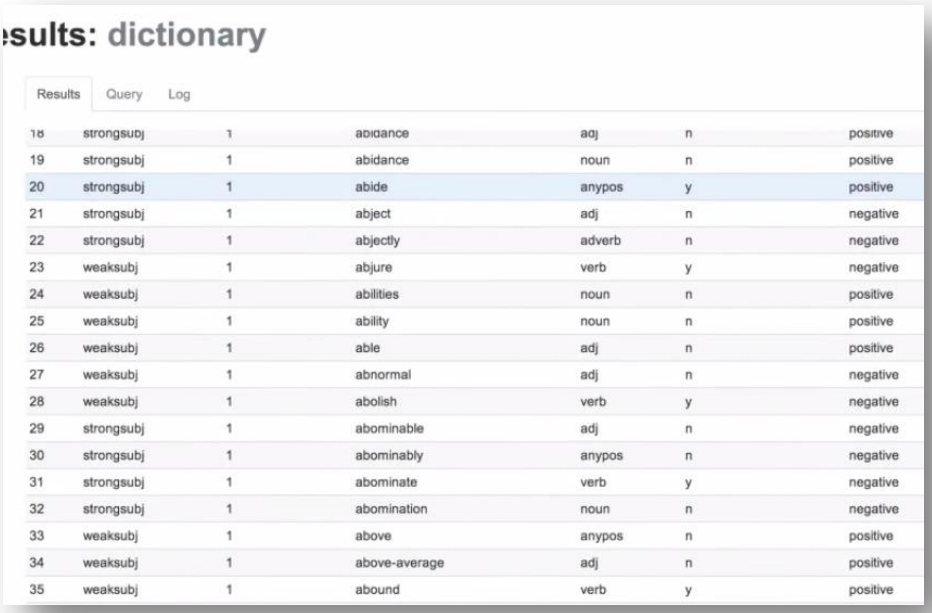
O script `hiveddl.sql` executou as seguintes etapas para refinar os dados:

Os dados do *Twitter* foram convertidos em um formato de tabela. (Figura 23):

Figura 23 - Tabela tweets_raw
Fonte: Autoria Própria.

3.4.3 Classificação de sentimento em positivo, negativo ou neutro.

A tabela `dictionary` foi usada para marcar o sentimento de cada *Tweet* pelo número de palavras positivas em comparação com o número de palavras negativas, e, em seguida, foi atribuído um valor de sentimento positivo, negativo ou neutro para cada *Tweet*. (Figura 24):



esults: dictionary

Results | Query | Log

18	strongsubj	1	abidance	adj	n	positive
19	strongsubj	1	abidance	noun	n	positive
20	strongsubj	1	abide	anypos	y	positive
21	strongsubj	1	abject	adj	n	negative
22	strongsubj	1	abjectly	adverb	n	negative
23	weaksubj	1	abjure	verb	y	negative
24	weaksubj	1	abilities	noun	n	positive
25	weaksubj	1	ability	noun	n	positive
26	weaksubj	1	able	adj	n	positive
27	weaksubj	1	abnormal	adj	n	negative
28	weaksubj	1	abolish	verb	y	negative
29	strongsubj	1	abominable	adj	n	negative
30	strongsubj	1	abominably	anypos	n	negative
31	strongsubj	1	abominate	verb	y	negative
32	strongsubj	1	abomination	noun	n	negative
33	weaksubj	1	above	anypos	n	positive
34	weaksubj	1	above-average	adj	n	positive
35	weaksubj	1	abound	verb	y	positive

Figura 24 - Tabela Dictionary
Fonte: Hortonworks Sandbox (2013).

Uma nova tabela foi criada incluindo o valor de sentimento para cada *Tweet.* Para isso foi feita uma divisão entre os *tweets* em três tipos de sentimento, foram atribuídos para comentários positivos o índice 2, para os comentários neutros 1 e para os comentários negativos índice 0 (Figura 25):

```
1  CREATE TABLE tweetsbi
2  STORED AS RCFile
3  AS
4  SELECT
5    t.*,
6    case s.sentiment
7      when 'positive' then 2
8      when 'neutral' then 1
9      when 'negative' then 0
10   end as sentiment
11 FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.id = s.id;
12
```

Figura 25 - Tabela Tweetsbi
Fonte: Hortonworks Sandbox (2013).

Podem-se analisar os dados usando a linha de comando *Hive.* Com o comando "`show tables`" pode-se ver as tabelas que foram adicionadas ao diretório (Figura 26):

```
hive> show tables;
OK
dictionary
l1
l2
l3
sample_07
sample_08
time_zone_map
tweets_clean
tweets_raw
tweets_sentiment
tweets_simple
tweetsbi
Time taken: 1.596 seconds, Fetched: 12 row(s)
hive>
```

Figura 26 - Listagem das tabelas adicionadas ao diretório Sandbox.
Fonte: Autoria Própria.

Na aba *HCatalog*, tabela é possível ver a listagem das tabelas geradas, para visualizar o conteúdo clica-se em `Browser Data`. (Figura 27):

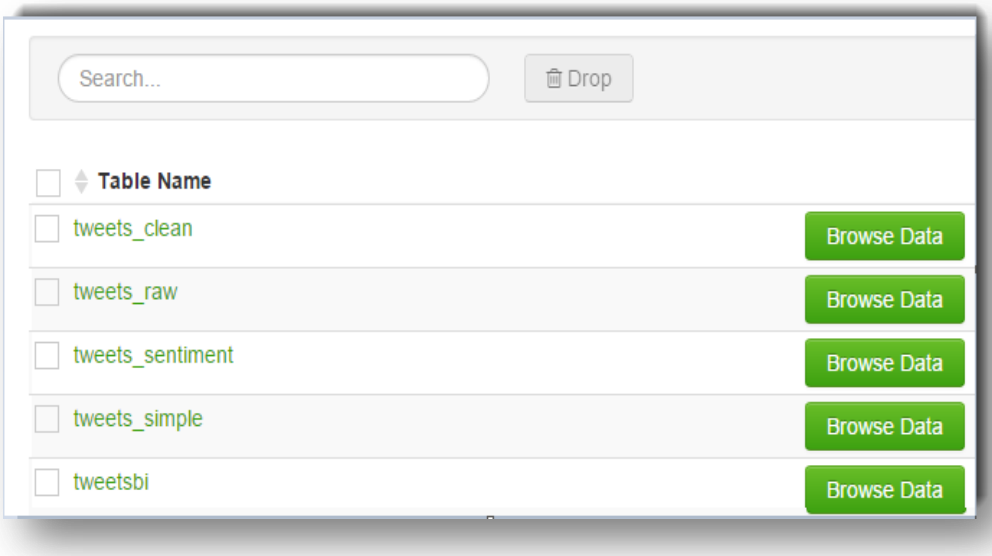


Figura 27 - Tabelas na aba HCatalog
Fonte: Autoria Própria.

Esta tabela foi criada pelo script *Hive* em que foi acrescentada uma coluna com o valor de sentimento para cada *tweet*.(Figura 28):

Figura 28 - Query Results
Fonte: Hortonworks,Inc (2013).

4 RESULTADOS E DISCUSSÃO

Uma vez que os dados já tenham sido transformados em formato tabular, os resultados podem ser verificados por meio de análise de sentimento utilizando a ferramenta Microsoft Excel.

4.1 IMPORTAÇÃO DOS DADOS NO EXCEL

Em uma nova planilha, selecionam-se `Dados > De Outras Fontes > Do Microsoft Query`, **(**Figura 29).

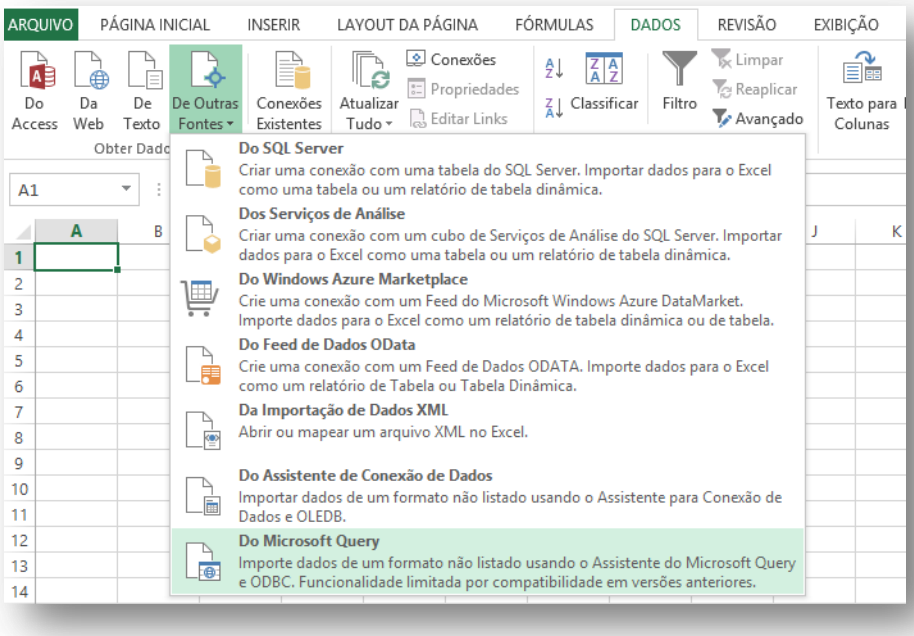


Figura 29 - Microsoft Query
Fonte: Autoria Própria.

Na Figura 30 tem-se a opção `Escolher Fonte de Dados`, em que se escolhe a fonte de dados ODBC *Hortonwork*s. O driver *Hortonworks* ODBC permite que se tenha acesso aos dados *Hortonworks* com outras aplicações de *Business Intelligence* (BI) que suportam ODBC Excel.

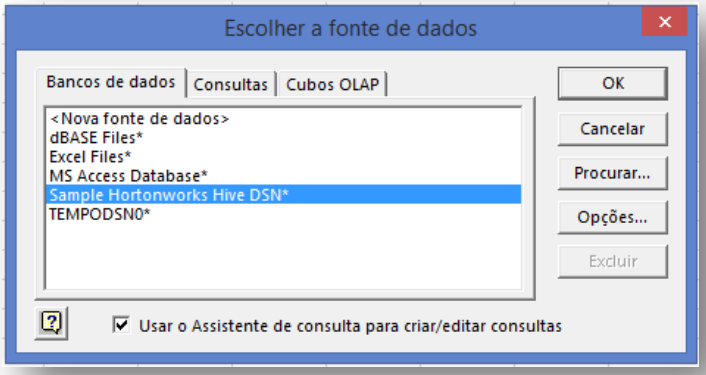


Figura 30 - Escolha da Fonte de Dados
Fonte: Autoria Própria.

Após a conexão com o *Sandbox* ser estabelecida, aparece o Assistente de Consulta. Seleciona-se a tabela "`tweetsbi`" nas tabelas e colunas disponíveis, em seguida, adicione toda a tabela "`tweetsbi`" para a consulta. (Figura 31):

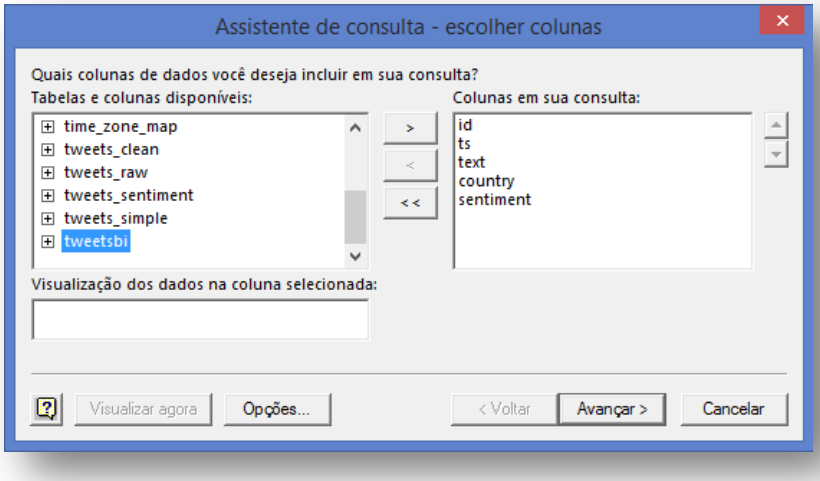
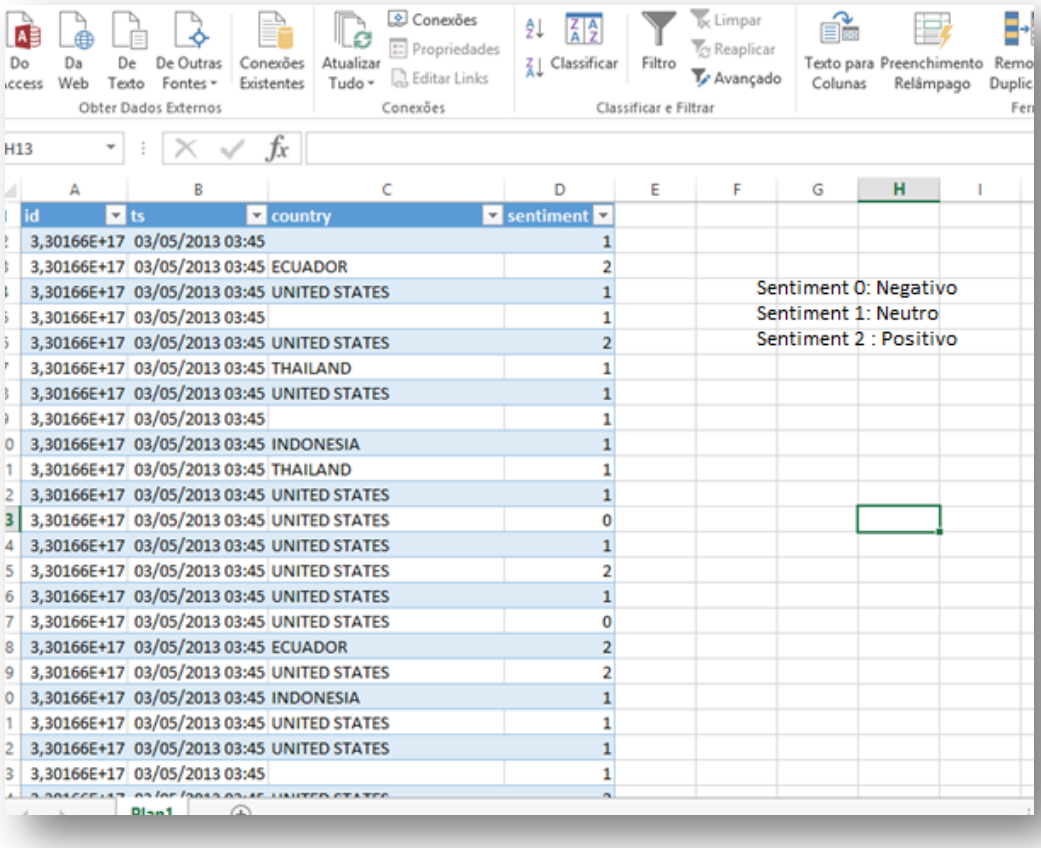


Figura 31 - Escolher colunas
Fonte: Autoria Própria.

Os dados importados aparecem em forma de tabela como mostra a Figura 32:



id	ts	country	sentiment
3,30166E+17	03/05/2013 03:45		1
3,30166E+17	03/05/2013 03:45	ECUADOR	2
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45		1
3,30166E+17	03/05/2013 03:45	UNITED STATES	2
3,30166E+17	03/05/2013 03:45	THAILAND	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45		1
3,30166E+17	03/05/2013 03:45	INDONESIA	1
3,30166E+17	03/05/2013 03:45	THAILAND	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	0
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	2
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	0
3,30166E+17	03/05/2013 03:45	ECUADOR	2
3,30166E+17	03/05/2013 03:45	UNITED STATES	2
3,30166E+17	03/05/2013 03:45	INDONESIA	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45	UNITED STATES	1
3,30166E+17	03/05/2013 03:45		1

Figura 32 - Dados Importados
Fonte: Autoria Própria.

4.2 VISUALIZAÇÃO DOS RESULTADOS

Após os dados de sentimento do *Twitter* forem importados com êxito para a Microsoft Excel, pode-se usar o recurso *Excel Power View* para analisar e visualizar os estes dados. (Figura 33):

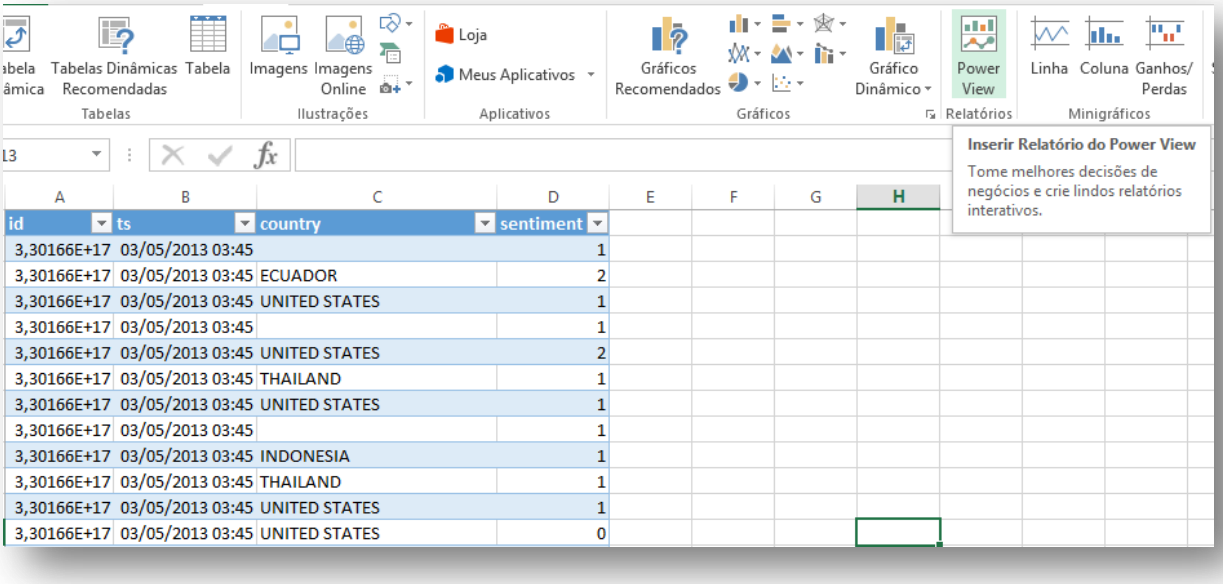


Figura 33 - Inserir Power View
Fonte: Autoria Própria.

Na área de `Campos Power View`, deve-se desmarcar as caixas de seleção ao lado dos campos `id` e `ts`, e clica-se em `Mapa` na guia `Design` no menu superior.(Figura 34):

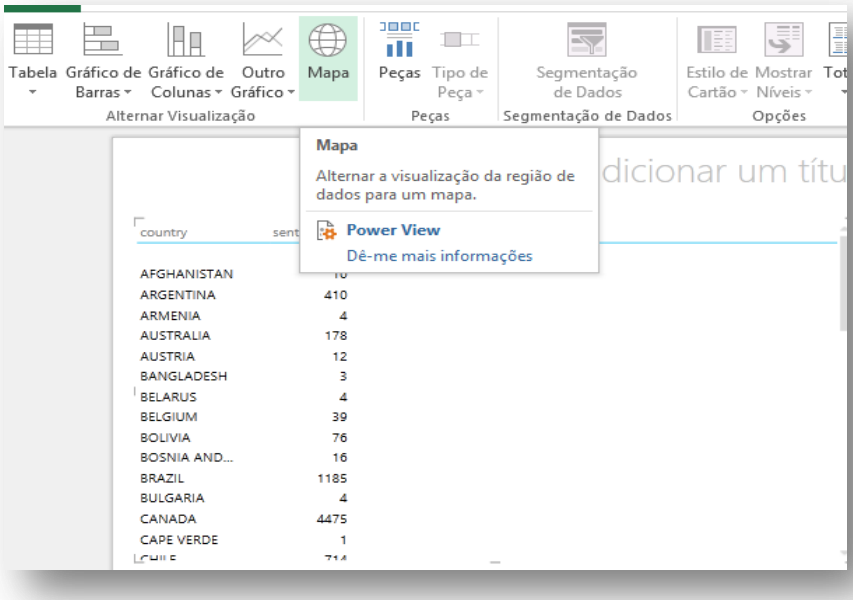


Figura 34- Visualizar dados com o Mapa
Fonte: Autoria Própria.

É possível ter uma visão global dos dados, (Figura 35). Estas bolinhas representam a quantidade de comentários feita pelas pessoas em cada país.

Figura 35 - Visualização do Mapa Global: Sentimento x País
Fonte: Autoria Própria.

Devem-se exibir os dados de sentimento pela cor. Comentários negativos são representados pela cor azul, comentários de sentimento neutro pela cor vermelha e comentários positivos pela cor laranja. (Figura 36):

Figura 36 - Contagem de tweets por país e sentimento
Fonte: Autoria Própria.

O mapa apresenta os dados de sentimento do Brasil. Existem 841 *tweets* tem uma pontuação neutra de sentimento, como indicado pela cor vermelha (Figura 37):

Figura 37 - Contagem de sentimentos neutros no Brasil
Fonte: Autoria Própria.

Na Figura 38, têm-se os Estados Unidos da América com 13.235 *tweets* de sentimento positivo indicado pela cor laranja.

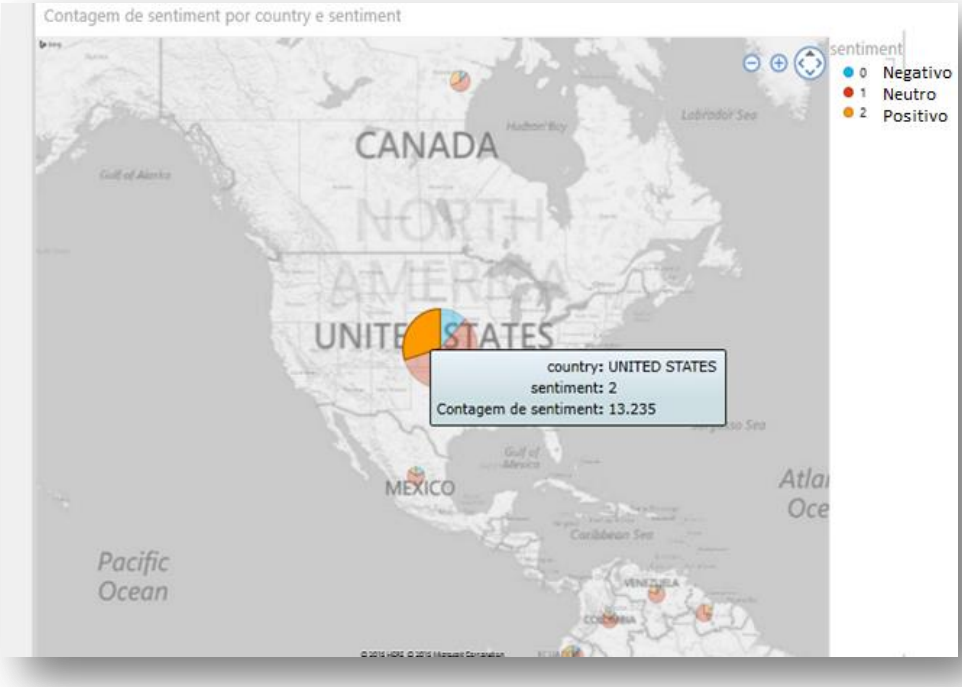


Figura 38 - Contagem de Sentimentos Positivos nos Estados Unidos
Fonte: Autoria Própria.

E também se tem o Canadá com uma contagem de 426 *tweets* negativos sobre o filme. (Figura 39).

Figura 39 - Contagem de sentimentos negativos no Canadá.
Fonte: Autoria Própria.

4.3 APLICAÇÃO DOS RESULTADOS

As organizações usam análise de sentimentos para entender como o público se sente sobre algo em um momento particular no tempo, e também para acompanhar como essas opiniões mudam com o tempo. (HORTONWORKS, 2013).

Uma empresa pode analisar sentimento sobre:

- Um produto - Por exemplo, faz o segmento-alvo compreender e apreciar mensagens em torno de um lançamento de produto. Quais os produtos que os visitantes tendem a comprar em conjunto, e quais são eles mais propensos a comprar no futuro?
- Um serviço - Por exemplo, um hotel ou restaurante tem um serviço de boa ou má qualidade;

- Concorrentes - Em que áreas é que as pessoas veem a empresa como melhor do que (ou mais fraco do que) a concorrência;
- Reputação - O que o público realmente pensa sobre a reputação da empresa;

Neste caso, a análise foi concentrada no lançamento de um produto, onde os dados relativos ao perfil das pessoas que estavam assistindo ou poderiam vir a assistir ao filme Iron Man, e em seguida estariam expressando suas opiniões com mensagens por meio do *Twitter*.

Por meio do Microsoft Excel foi possível ter uma visualização geral de todos os dados de sentimento do *Twitter* gerados pelas ferramentas anteriormente citadas, e a realização da análise das opiniões de todas as pessoas que assistiram ao filme e fizeram comentários positivos ou negativos em diversos locais do mundo. Estas informações puderam ser mais bem analisadas com a utilização da ferramenta de Mapa do *Excel Power View*, dividindo os tipos de sentimentos por cores, facilitando assim a análise e a contagem destes *tweets.*

Neste estudo, os dados gerados poderão ser úteis para a melhor compreensão das empresas sobre o que seus clientes pensam, tendo assim um melhor planejamento de suas atividades de marketing para possíveis futuros lançamentos de novos produtos.

5 CONSIDERAÇÕES FINAIS

5.1 CONCLUSÃO

Após a realização deste trabalho e análise dos seus resultados é possível verificar que:

a) Dentre as diversas tecnologias presentes em *Big Data*, a *text mining* é responsável pela análise destes dados não estruturados. A análise de textos livres possui uma complexidade superior à análise de dados estruturados e o seu principal objetivo é encontrar padrões para conseguir extrair informações destes textos, podendo assim fazer uma análise de sentimento destes textos.
b) Com a utilização da plataforma *Hortonworks Data Plataform* é possível interagir com outros componentes de dados dentro do ecossistema *Hadoop e Business Intelligente*. A HDP se integra e aumenta seus aplicativos e sistemas existentes para que se aproveite o *Hadoop* com mudanças mínimas nos conjuntos de habilidades e arquiteturas de dados existentes.
c) Com a utilização do *Apache Hadoop* é possível fazer uso da análise de enormes quantidades de dados criados pela opinião dos clientes sobre determinados produtos nas redes sociais, podendo se obter uma visão geral do cliente, englobando mídias sociais, sequência de cliques, vídeo e dados de transações sem a necessidade de um esquema de dados predefinido. Obtendo assim um retorno real do consumidor em relação ao atendimento, a qualidade do produto ou o impacto que uma propaganda de marketing tem sobre os usuários.
d) A análise de sentimentos é uma área de crescente interesse. Neste estudo, foram discutidos seus conceitos básicos, técnicas que podem ser usadas para identificar, classificar a polaridade e agregar o sentimento expresso. O volume crescente de conteúdo subjetivo disponível diariamente, em particular nas redes sociais, motiva o crescimento da área com novas técnicas capazes de processar automaticamente textos. Muitas são as aplicações centradas na visualização do sentimento, ou na predição de comportamentos com base no sentimento existente.

e) Por fim, apesar dos desafios, a possibilidade de trabalhar com amostras maiores de dados das redes sociais permite que novas informações sejam extraídas e que informações sejam mais consistentes, visto que a amostra analisada será maior. Tratar essas novas informações adequadamente extrapola as áreas técnicas da computação, tendo em vista que conhecimentos de áreas de humanas (como antropologia, sociologia, psicologia, entre outros) são necessários. Este trabalho introduziu o tema de *Big Data* e análise de sentimentos nas redes sociais, permitindo que pesquisadores e analistas que desejam trabalhar com grande volume de dados conheçam as principais abordagens e desafios que existem atualmente.

5.2 TRABALHOS FUTUROS/CONTINUAÇÃO DO TRABALHO

Propõe-se como trabalho futuro, o estudo mais detalhado e aplicação da ferramenta *Apache Flume* na obtenção dos dados brutos advindos do *Twitter*, sendo utilizado como complemento do atual estudo.

Também se tem como possibilidade o uso da análise de sentimentos para aplicação de um serviço, como exemplo um hotel ou restaurante poder analisar a opinião de seus clientes nos locais em que prestam serviço. Ou também sobre a opinião dos clientes sobre a reputação de uma determinada empresa.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BERNARD, Allen. ***Big Data*** **valoriza o Business Intelligence.** Disponível em:<http://cio.com.br/tecnologia/2012/10/23/big-data-valoriza-o-business-intelligence/> Acesso em 20 set. 2014

BRANCO, Tom. **Hadoop: The Definitive Guide** . O'ReillyMedia . ISBN 978-1-4493-8973-4 . ago.2010.

BROWN, B; CHUI, M; MANYIKA, J. **Are you ready for the era of '*Big Data*'?.** McKinsey Global Institute**.** 2011

BUGHIN, J; LIVINGSTON, J; Marwaha, S. Seizing **thepotentialof 'bigdata**'.McKinsey Global Institute. 2011

BRYNJOLFSSON, E; MCAFEE, A. ***Big Data*** **- A Revolução da Gestão**. HarvarBusiness Review**.** 2012

CARDOZO, M. L.. **Twitter: microblog e rede social**. Caderno.Com. São Caetano do Sul, v.4, n.2, p.24-38 , 2ºsemestre 2009.

CEARLEY, D; Claunch, C. The **Top 10 Strategic Technology Trends** for 2013. Gartner**.** 2013

CHUCK, Lam. **Hadoop emação.**Manning Publications.ISBN –1935182196, 2009.

CISCO. ***Big Data*****: Grande volume de Dados, Grande Potencial, Grande Prioridade** Disponível em: http://www.cisco.com/*web*/PT/press/articles/2013/20130401.html Acesso em: 18 Out.2014.

COMPUTER WORLD. ***Big Data*** **ajuda grandes empresas a incrementar receita**.. Disponível em: http://computerworld.uol.com.br/gestao/2013/06/17/big-data-ajuda-grandes-empresasaincrementar-receita-aponta-estudo/ . Acesso em: 18 out.2014.

COSTA, Luís Henrique M. K.; AMORIM, Marcelo D. de; CAMPISTA, Miguel Elias M.; RUBINSTEIN, Marcelo G.; FLORISSI, Patrícia; DUARTE, Otto Carlos M. B. **Grandes massas de dados na nuvem: desafios e técnicas para inovação.** In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES SBRC, 2012, Ouro Preto, Minas Gerais, maio, 2012.

COSTA, L. et al. **Redes: uma introdução às dinâmicas da conectividade e da auto-organização**. Brasília: WWF-Brasil, 2003.Statistics | Facebook (2011), Disponível em: http://www.facebook.com/press/info.php?statistics, Acesso em: 20 jun.2014.

DAVENPORT, T; BARTH, P; BEAN, R. How'***Big Data***' **isDifferent**. MIT SloanManagement Review**.** 2012

ELIAS, Diego, **A diferença entre *Big Data* e Business Intelligence**. Disponível em:<http://corporate.canaltech.com.br/noticia/business-intelligence/A-diferenca-entre-Big-Data-e-Business-Intelligence/>. Acesso em 30 set.2014.

ELIAS,Diego. **Contextualizando *Big Data*.** Disponível em: <http://www.binapratica.com.br/#!contextualizando-big-data/c1pur>. Acesso em 01 outubro de 2014.

EMC. **Homepage da Instituição**. Disponível em: <http://brazil.emc.com/collateral/emcperspective/h8729-gain-comp-advtng-ep.pdf>. Acesso em: 18 Jun.2013.

ESTADÃO. **Como o facebook rastreia os usuários**. Disponível em: http://blogs.estadao.com.br/link/como-o-facebook-rastreia-os-usuarios/ . Acesso em: 18 out.2014.

FEIJÓ, Bruno Vieira..**Revista Exame PME – Pequenas e Médias Empresas:** A Revolução dos Dados. São Paulo, p. 30-43, set. 2013.

GALLANT, J. TIBCO CEO: **How Real-Time Computing Will Change the Landscape.**ComputerWorld.2011

GANTZ, John; REINSEL, David. **Extractingvaluefromchaos**. Disponível em: <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>. Acesso em: 13 ago. 2014.

GANTZ, J; REINSEL, D. THE DIGITAL UNIVERSE IN 2020: ***Big Data*, BiggerDigital Shadows, andBiggestGrowth in theFarEast**. IDC. 2012

GASPAROTTO, Henrique. **HadoopMapReduce**: Introdução a *Big Data*. Disponivel em:<http://www.devmedia.com.br/hadoop-mapreduce-introducao-a-big-data/30034>. Acesso em 30 set.2014.

GOLDMAN, Alfredo; KON, FABIO; JUNIOR, Francisco; POLATO, Ivanilton; PEREIRA, Rosangela. **Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades. Disponivel em:** http://www.ime.usp.br/~ipolato/JAI2012-Hadoop.pdf. Acesso em 24 mar.2014.

HANSON,J.Jeffey**. Uma Introdução ao HadoopDistributed File System.** Disponível em:<http://www.ibm.com/developerworks/br/library/wa-introhdfs/>. Acesso em 20 set 2014.

HEA, H; ZHAB, S; LI, L. **Social media competitiveanalysisandtext mining: A case study in the pizza industry**. InternationalJournalofInformation Management**.** 2013

IBM,**Emeryville**.McGraw-Hill Osborne Media,2012.Disponivel em: <http://public.dhe.ibm.com/common/ssi/ecm/en/imm14100usen/IMM14100USEN.PDF>Acesso em: 15 ago 2014.

IBM. ***Big Data* trazendo novas oportunidades e vantagens competitivas**. Disponível em: <https://www.ibm.com/developerworks/community/blogs/fd26864dcb4149cfb719d89c6b072893/entry/big_data_trazendo_novas_oportunidades_e_vantagens_competitivas?lang=pt_br> Acesso em: 18 out.2014.

INFO,ABRIL(2015). Avaliação Info, Win SCP. Disponível em: http://info.abril.com.br/downloads/windows/winscp. Acessado em 18 Mai. 2015.

ISACA®, ***Big Data* – Impactos e Benefícios**. 2013. Disponível em: <http://www.isaca.org/Knowledge-Center/Research/Documents/Big-Data_whp_Por_0413.pdf?regnum=218726> Acesso em: 14 ago. 2014.

MACHADO,André.**Estudo da EMC prevê que volume de dados virtuais armazenados será seis vezes maior em 2020**. O GLOBO, 2014. Disponível em:http://oglobo.globo.com/sociedade/tecnologia/estudo-da-emc-preve-que-volume-de-dados-virtuais-armazenados-sera-seis-vezes-maior-em-2020-12147682>. Acessoem 10 set. 2014.

MANOVICH, L. (2011), "**Trending: the promises and the challenges of big social data**",Minneapolis, MN: University of Minnesota Press.

MANYIKA, James; CHUI, Michael; BROWN**, Brad; et al. *Big Data*:** The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.

MASSUDA,Felipe. **Indrodução ao Hadoop.** Disponível em:<http://www.aqueleblogdesoa.com.br/2013/05/introducao-ao-hadoop/>.Acesso em 30 set.2014.

OLIVEIRA, Vinicius. **Tudo sobre Business Intelligence** Disponível em: http://www.binapratica.com.br/#!visao-pentaho/c7gy . Acesso em 13 mar. 2015.

PAL, Kaushik.**Whats is the difference between *Big Data* and *Data Mining*.**Disponível em:<http://www.techopedia.com/7/29678/technology-trends/what-is-thedifferencbetween-big-data-and-data-mining>. Acesso em 29 set.2014.

PETRY, André; VILICIC, Filipe. **A era do *Big Data* e dos algoritmos está mudando o mundo.**RevistaVeja. n.20, maio, 2013.p.70-81.

PROKOPP, Christian. **The Free Hive Book.**Disponivel em: http://www.semantikoz.com/blog/the-free-apache-hive-book/. Acesso em 30 set.2014.

RUBINSTEIN, Marcelo G.; FLORISSI, Patrícia; DUARTE, Otto Carlos M. B. **Grandes massas de dados na nuvem: desafios e técnicas para inovação.** In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES SBRC, 2012, Ouro Preto, Minas Gerais, maio, 2012.

SAS. ***Big Data*.** Disponível em: <http://www.sas.com/offices/latinamerica/brazil/news/preleases/pesquisa-sasbigdata. HTML> Acesso em: 18 out.2014.

SHEPS, Swain. ***Business inteligence*For Dummies** .Wiley Publishing Inc, 2013
State of The World. Genebra: World Economic Forum, 2012.

SOLIS, B. (2007), "**Manifesto, The Social Media**", Disponível em: http://www.briansolis.com/2007/06/future-of-communications-manifesto-for/ , Acesso em: 20 jun 2014.

SULLIVAN, DAN.(2014). **Getting Started with Hadoop 2.0** Disponivel em: http://www.tomsitpro.com/articles/hadoop-2-vs-1,2-718.html. Acessado em 26 Dez. 2014.

TAUBE, B. Leveraging ***Big Data*** **and realtime analytics to achieve situational awareness for smart grids (white paper)**.Versant Corporation U.S.Headquarters**,** Redwood City 2012.

TECHAMERICA. TechAmericaFoundation Federal *Big Data* Commision. **Demystifying *Big Data*: A PracticalGuidetoTransforming The Business of Government**, 2012. Disponível em: <http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareportfinal.pdf>, Acesso em: 14 Out. 2014.

VENNER, Jason .**Pro Hadoop**. Apress, ISBN 978-1-4302-1942-2.,set. 2009.

WEF (World Economic Forum). **Listamantidapelo Committed to Improviment, 2012.**

WEF. World EconomicForum. ***Big Data*, Big Impact: New Possibilities for InternationalDevelopment**, 2012. Disponível em <http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf>, Acesso em 21 Out. 2014.

WHITE, Tom. **Hadoop: The Definitive Guide**. 2 ed. Cambridge: O'Reilly, 2010.

XIAO Zhifeng, XIAO Yang. Achieving **Accountable MapReduce in *Cloud Computing*.** Disponível em:<http://www.sciencedirect.com/science/article/pii/S0167739X13001465>. Acesso em 10 set.2014.

ZIKOPOULOS, P; DE ROOS, D; PARASURAMAN, K; DEUTSCH, T; GILES, J;CORRIGAN, D. **Harness the power of *Big Data*- The IBM *Big Data* Platform.**Emeryville: McGraw-Hill Osborne Media, 2012.Disponivel em: <http://public.dhe.ibm.com/common/ssi/ecm/en/imm14100usen/IMM14100USEN.PDF>Acesso em Ago.2014.